

Low-rank Matrix Completion using Alternating Minimization

Prateek Jain
Microsoft Research India, Bangalore
prajain@microsoft.com

Praneeth Netrapalli
The University of Texas at Austin
praneethn@utexas.edu

Sujay Sanghavi
The University of Texas at Austin
sanghavi@mail.utexas.edu

December 4, 2012

Abstract

Alternating minimization represents a widely applicable and empirically successful approach for finding low-rank matrices that best fit the given data. For example, for the problem of low-rank matrix completion, this method is believed to be one of the most accurate and efficient, and formed a major component of the winning entry in the Netflix Challenge [15, 16].

In the alternating minimization approach, the low-rank target matrix is written in a *bi-linear form*, i.e. $X = UV^\dagger$; the algorithm then alternates between finding the best U and the best V . Typically, each alternating step in isolation is convex and tractable. However the overall problem becomes non-convex and there has been almost no theoretical understanding of when this approach yields a good result.

In this paper we present first theoretical analysis of the performance of alternating minimization for matrix completion, and the related problem of matrix sensing. For both these problems, celebrated recent results have shown that they become well-posed and tractable once certain (now standard) conditions are imposed on the problem. We show that alternating minimization *also* succeeds under similar conditions. Moreover, compared to existing results, our paper shows that alternating minimization guarantees faster (in particular, geometric) convergence to the true matrix, while allowing a simpler analysis.

1 Introduction

Finding a low-rank matrix to fit / approximate observations is a fundamental task in data analysis. In a slew of applications, a popular empirical approach has been to represent the target rank k matrix $X \in \mathbb{R}^{m \times n}$ in a *bi-linear form* $X = UV^\dagger$, where $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$. Typically, this is done for two reasons:

- (a) *Size and computation:* If the rank k of the target matrix (to be estimated) is much smaller than m, n , then U, V are significantly smaller than X and hence are more efficient to optimize for. This is crucial for several practical applications, e.g., recommender systems where one routinely encounters matrices with billions of entries.
- (b) *Modeling:* In several applications, one would like to impose extra constraints on the target matrix, besides just low rank. Oftentimes, these constraints might be easier and more natural to

impose on factors U, V . For example, in Sparse PCA [22], one looks for a low-rank X that is the product of *sparse* U and V .

Due to the above two reasons, in several applications, the target matrix X is parameterized by $X = UV^\dagger$. For example, clustering [14], sparse PCA [22] etc.

Using the bi-linear parametrization of the target matrix X , the task of estimating X now reduces to finding U and V that, for example, minimize an error metric. The resulting problem is typically non-convex due to bi-linearity. Correspondingly, a popular approach has been to use **alternating minimization**: iteratively keep one of U, V fixed and optimize over the other, then switch and repeat, see e.g. [16]. While the overall problem is non-convex, each sub-problem is typically convex and can be solved efficiently.

Despite wide usage of bi-linear representation and alternating minimization, there has been to date almost no theoretical understanding of when such a formulation works. Motivated by this disconnect between theory and practice in the estimation of low-rank matrices, in this paper we provide the **first guarantees for performance of alternating minimization**, for two low-rank matrix recovery problems: matrix completion, and matrix sensing.

Matrix completion involves completing a low-rank matrix, by observing only a few of its elements. Its recent popularity, and primary motivation, comes from recommendation systems [16], where the task is to complete a user-item ratings matrix using only a small number of ratings. As elaborated in Section 2, alternating minimization becomes particularly appealing for this problem as it provides a fast, distributed algorithm that can exploit both sparsity of ratings as well as the low-rank bi-linear parametrization of X .

Matrix sensing refers to the problem of recovering a low-rank matrix $M \in \mathbb{R}^{m \times n}$ from affine equations. That is, given d linear measurements $b_i = \text{tr}(A_i^\dagger M)$ and measurement matrices A_i 's, the goal is to recover back M . This problem is particularly interesting in the case of $d \ll mn$ and was first studied in [19] and subsequently in [10, 17]. In fact, matrix completion is a special case of this problem, where each observed entry in the matrix completion problem represents one single-element measurement matrix A_i .

Without any extra conditions, both matrix sensing and matrix completion are ill-posed problems, with potentially multiple low-rank solutions, and are in general NP hard [18]. Current work on these problems thus impose some extra conditions, which makes the problems both well defined, and amenable to solution via the respective proposed algorithms [19, 3]; see Section 3 for more details. In this paper, we show that **under similar conditions to the ones used by the existing methods**, alternating minimization also guarantees recovery of the true matrix; we also show that it requires only a small number of computationally cheap iterations and hence, as observed empirically, is computationally much more efficient than the existing methods.

Notations: We represent a matrix by capital letter (e.g. M) and a vector by small letter (u). u_i represents i -th element of u and U_{ij} denotes (i, j) -th entry of U . U_i represents i -th column of U and $U^{(i)}$ represents i -th row of U . A^\dagger denotes matrix transpose of A . $u = \text{vec}(U)$ represents vectorized U , i.e., $u = [U_1^\dagger U_2^\dagger \dots U_k^\dagger]^\dagger$. $\|u\|_p$ denotes L_p norm of u , i.e., $\|u\|_p = (\sum_i |u_i|^p)^{1/p}$. By default, $\|u\|$ denotes L_2 norm of u . $\|A\|_F$ denotes Frobenius norm of A , i.e., $\|\text{vec}(A)\|_2$. $\|A\|_2 = \max_{x, \|x\|_2=1} \|Ax\|_2$ denotes spectral norm of A . $\text{tr}(A)$ denotes the trace (sum of diagonal elements) of square matrix A . Typically, \hat{U}, \hat{V} represent factor matrices (i.e., $\hat{U} \in \mathbb{R}^{m \times k}$ and $\hat{V} \in \mathbb{R}^{n \times k}$) and U, V represent their orthonormal basis.

2 Our Results

In this section, we will first define the matrix sensing problem, and present our results for it. Subsequently, we will do the same for matrix completion. The matrix sensing setting – i.e. recovery of any low-rank matrix from linear measurements that satisfy matrix RIP – represents an easier analytical setting than matrix completion, but still captures several key properties of the problem that helps us in developing an analysis for matrix completion. We note that for either problem, ours represent the first global optimality guarantees for alternating minimization based algorithms.

Matrix Sensing via Alternating Minimization

Given d linear measurements $b_i = \langle M, A_i \rangle = \text{tr}(A_i^\dagger M)$, $1 \leq i \leq d$ of an *unknown* rank- k matrix $M \in \mathbb{R}^{m \times n}$ and the sensing matrices A_i , $1 \leq i \leq d$, the goal in matrix sensing is to recover back M . In the following we collate these coefficients, so that $b \in \mathbb{R}^d$ is the vector of b_i 's, and $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^d$ is the corresponding linear map, with $b = \mathcal{A}(M)$. With this notation, the Low-Rank Matrix Sensing problem is:

$$\text{Find } X \in \mathbb{R}^{m \times n}, \text{ s.t. } \mathcal{A}(X) = b, \text{ rank}(X) \leq k. \quad (\text{LRMS})$$

As in the existing work [19] on this problem, we are interested in the **under-determined case**, where $d < mn$. Note that this problem is a strict generalization of the popular compressed sensing problem [4]; compressed sensing represents the case when M is restricted to be a diagonal matrix.

For matrix sensing, alternating minimization approach involves representing X as a product of two matrices $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$, i.e., $X = UV^\dagger$. If k is (much) smaller than m, n , these matrices will be (much) smaller than X . With this bi-linear representation, alternating minimization can be viewed as an approximate way to solve the following non-convex optimization problem:

$$\min_{U \in \mathbb{R}^{m \times k}, V \in \mathbb{R}^{n \times k}} \| \mathcal{A}(UV^\dagger) - b \|_2^2$$

As mentioned earlier, alternating minimization algorithm for matrix sensing now alternately solves for U and V while fixing the other factor. See Algorithm 1 for a pseudo-code of AltMinSense algorithm that we analyze.

We note two key properties of AltMinSense : a) Each minimization – over U with V fixed, and vice versa – is a simple least-squares problem, which can be solved in time $O(dn^2k^2 + n^3k^3)$ ¹, b) We initialize U^0 to be the top- k left singular vectors of $\sum_i A_i b_i$ (step 2 of Algorithm 1). As we will see later in Section 4, this provides a good initialization point for the sensing problem which is crucial; if the first iterate \hat{U}^0 is orthogonal, or almost orthogonal, to the true U^* subspace, AltMinSense may never converge to the true space (this is easy to see in the simplest case, when the map is identity, i.e. $\mathcal{A}(X) = X$ – in which case AltMinSense just becomes the power method).

In general, since $d < mn$, problem (LRMS) is not well posed as there can be multiple rank- k solutions that satisfy $\mathcal{A}(X) = b$. However, inspired by a similar condition in compressed sensing [4], Recht et al. [19] showed that if the linear map \mathcal{A} satisfies a (*matrix*) *restricted isometry property* (RIP), then a trace-norm based convex relaxation of (LRMS) leads to exact recovery. This property is defined below.

¹Throughout this paper, we assume $m \leq n$.

Algorithm 1 AltMinSense : Alternating minimization for matrix sensing

- 1: Input b, \mathcal{A}
 - 2: Initialize \hat{U}^0 to be the top- k left singular vectors of $\sum_i A_i b_i$
 - 3: **for** $t = 0, \dots, T - 1$ **do**
 - 4: $\hat{V}^{t+1} \leftarrow \operatorname{argmin}_{V \in \mathbb{R}^{n \times k}} \|\mathcal{A}(\hat{U}^t V^\dagger) - b\|_2^2$
 - 5: $\hat{U}^{t+1} \leftarrow \operatorname{argmin}_{U \in \mathbb{R}^{m \times k}} \|\mathcal{A}(U (\hat{V}^{t+1})^\dagger) - b\|_2^2$
 - 6: **end for**
 - 7: Return $X = \hat{U}^T (\hat{V}^T)^\dagger$
-

Definition 2.1. [19] A linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^d$ is said to satisfy k -RIP, with δ_k RIP constant, if for all $X \in \mathbb{R}^{m \times n}$ s.t. $\operatorname{rank}(X) \leq k$, the following holds:

$$(1 - \delta_k) \|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1 + \delta_k) \|X\|_F^2. \quad (1)$$

Several random matrix ensembles with sufficiently many measurements (d) satisfy matrix RIP [19]. For example, if $d = \Omega(\frac{1}{\delta_k^2} kn \log n)$ and each entry of A_i is sampled i.i.d. from a 0-mean sub-Gaussian distribution then k -RIP is satisfied with RIP constant δ_k .

We now present our main result for AltMinSense.

Theorem 2.2. Let $M = U^* \Sigma^* V^{*\dagger}$ be a rank- k matrix with non zero singular values $\sigma_1^* \geq \sigma_2^* \dots \geq \sigma_k^*$. Also, let the linear measurement operator $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^d$ satisfy $2k$ -RIP with RIP constant $\delta_{2k} < \frac{(\sigma_k^*)^2}{(\sigma_1^*)^2} \frac{1}{100k}$. Then, in the AltMinSense algorithm (Algorithm 1), for all $T > 2 \log(\|M\|_F/\epsilon)$, the iterates \hat{U}^T and \hat{V}^T satisfy:

$$\|M - \hat{U}^T (\hat{V}^T)^\dagger\|_F \leq \epsilon.$$

The above theorem establishes geometric convergence (in $O(\log(1/\epsilon))$ steps) of AltMinSense to the optimal solution of (LRMS) under standard RIP assumptions. This is in contrast to existing iterative methods for trace-norm minimization all of which require at least $O(\frac{1}{\sqrt{\epsilon}})$ steps; interior point methods for trace-norm minimization converge to the optimum in $O(\log(1/\epsilon))$ steps but require storage of the full $m \times n$ matrix and require $O(n^5)$ time per step, which makes it infeasible for even moderate sized problems.

Recently, several projected gradient based methods have been developed for matrix sensing [10, 17] that also guarantee convergence to the optimum in $O(\log(1/\epsilon))$ steps. But each iteration in these algorithms requires computation of the top k singular components of an $m \times n$ matrix, which is typically significantly slower than solving a least squares problem (as required by each iteration of AltMinSense).

Stagewise AltMinSense Algorithm: A drawback of our analysis for AltMinSense is the dependence of δ_{2k} on the condition number ($\kappa = \frac{\sigma_1^*}{\sigma_k^*}$) of M , which implies that the number of measurements d required by AltMinSense grows quadratically with κ . We address this issue by using a stagewise version of AltMinSense (Algorithm 3) for which we are able to obtain near optimal measurement requirement.

The key idea behind our stagewise algorithm is that if one of the singular vectors of M is very dominant, then we can treat the underlying matrix as a rank-1 matrix plus noise and approximately recover the top singular vector. Once we remove this singular vector from the measurements, we will have a relatively well-conditioned problem. Hence, at each stage of Algorithm 3, we seek to

remove the remaining most dominant singular vector of M . See Section 6 for more details; here we state the corresponding theorem regarding the performance of Stage-AltMin.

Theorem 2.3. *Let $M = U^* \Sigma^* V^{*\dagger}$ be a rank- k incoherent matrix with non zero singular values $\sigma_1^* \geq \sigma_2^* \cdots \geq \sigma_k^*$. Also, let $\mathcal{A}(\cdot) : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^d$ be a linear measurement operator that satisfies $2k$ -RIP with RIP constant $\delta_{2k} < \frac{1}{3200k^2}$. Suppose, Stage-AltMin (Algorithm 3) is supplied inputs \mathcal{A} , $b = \mathcal{A}(M)$. Then, the i -th stage iterates $\hat{U}_{1:i}^T$, $V_{1:i}^T$ satisfy:*

$$\|M - \hat{U}_{1:i}^T (V_{1:i}^T)^\dagger\|_F^2 \leq \max(\epsilon, 16k(\sigma_{i+1}^*)^2),$$

where $T = \Omega(\log(\|M\|_F^2/\epsilon))$. That is, the T -th step iterates of the k -th stage, satisfy: $\|M - \hat{U}_{1:k}^T (V_{1:k}^T)^\dagger\|_F^2 \leq \epsilon$.

The above theorem guarantees exact recovery using $O(k^4 n \log n)$ measurements which is only $O(k^3)$ worse than the information theoretic lower bound. We also note that for simplicity of analysis, we did not optimize the constant factors in δ_{2k} .

Matrix Completion via Alternating Minimization

The matrix completion problem is the following: there is an unknown rank- k matrix $M \in \mathbb{R}^{m \times n}$, of which we know a set $\Omega \subset [m] \times [n]$ of elements; that is, we know the values of elements M_{ij} , for $(i, j) \in \Omega$. The task is to recover M . Formally, the Low-Rank Matrix Completion problem is:

$$\text{Find rank-}k \text{ matrix } X \text{ s.t. } P_\Omega(X) = P_\Omega(M), \quad (\text{LRMC})$$

where for any matrix S and a set of elements $\Omega \subset [m] \times [n]$ the matrix $P_\Omega(S) \in \mathbb{R}^{m \times n}$ is as defined below:

$$P_\Omega(S)_{ij} = \begin{cases} S_{ij} & \text{if } (i, j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

We are again interested in the under-determined case; in fact, for a fixed rank k , as few as $O(n \log n)$ elements may be observed. This problem is a special case of matrix sensing, with the measurement matrices $A_i = e_j e_\ell^\dagger$ being non-zero only in single elements; however, such matrices do not satisfy matrix RIP conditions like (1). For example, consider a low-rank $M = e_1 e_1^\dagger$ for which a uniformly random Ω of size $O(n \log n)$ will most likely miss the non-zero entry of M .

Nevertheless, like matrix sensing, matrix completion has been shown to be possible once additional conditions are applied to the low-rank matrix M and the observation set Ω . Starting with the first work [3], the typical assumption has been to have Ω generated uniformly at random, and M to satisfy a particular incoherence property that, loosely speaking, makes it very far from a sparse matrix. In this paper, we show that *once* such assumptions are made, alternating minimization *also* succeeds. We now restate, and subsequently use, this incoherence definition.

Definition 2.4. [3] *A matrix $M \in \mathbb{R}^{m \times n}$ is incoherent with parameter μ if:*

$$\|u^{(i)}\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{m}} \quad \forall i \in [m], \quad \|v^{(j)}\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{n}} \quad \forall j \in [n], \quad (3)$$

where $M = U \Sigma V^T$ is the SVD of M and $u^{(i)}$, $v^{(j)}$ denote the i^{th} row of U and the j^{th} row of V respectively.

The alternating minimization algorithm can be viewed as an approximate way to solve the following non-convex problem:

$$\min_{U, V \in \mathbb{R}^{n \times k}} \|P_\Omega(UV^\dagger) - P_\Omega(M)\|_F^2$$

Similar to AltMinSense, the altmin procedure proceeds by alternatively solving for U and V . As noted earlier, this approach has been popular in practice and has seen several variants and extensions being used in practice [21, 16, 15, 7]. However, for ease of analysis, our algorithm further modifies the standard alternating minimization method. In particular, we introduce partitioning of the observed set Ω , so that we use different partitions of Ω in each iteration. See Algorithm 2 for a pseudo-code of our variant of the alternating minimization approach.

Algorithm 2 AltMinComplete: Alternating minimization for matrix completion

- 1: Input: observed set Ω , values $P_\Omega(M)$
 - 2: Partition Ω into $2T + 1$ subsets $\Omega_0, \dots, \Omega_{2T}$ with each element of Ω belonging to one of the Ω_t with equal probability (sampling with replacement)
 - 3: $\hat{U}^0 = \text{SVD}(\frac{1}{p}P_{\Omega_0}(M), k)$ i.e., top- k left singular vectors of $\frac{1}{p}P_{\Omega_0}(M)$
 - 4: Clipping step : Set all elements of \hat{U}^0 that have magnitude greater than $\frac{2\mu\sqrt{k}}{\sqrt{n}}$ to zero and orthonormalize the columns of \hat{U}^0
 - 5: **for** $t = 0, \dots, T - 1$ **do**
 - 6: $\hat{V}^{t+1} \leftarrow \text{argmin}_{V \in \mathbb{R}^{n \times k}} \|P_{\Omega_{t+1}}(\hat{U}^t V^\dagger - M)\|_F^2$
 - 7: $\hat{U}^{t+1} \leftarrow \text{argmin}_{U \in \mathbb{R}^{m \times k}} \|P_{\Omega_{T+t+1}}(U(\hat{V}^{t+1})^\dagger - M)\|_F^2$
 - 8: **end for**
 - 9: Return $X = \hat{U}^T(\hat{V}^T)^\dagger$
-

We now present our main result for (LRMC):

Theorem 2.5. *Let $M = U^* \Sigma^* V^{*\dagger} \in \mathbb{R}^{m \times n}$ ($n \geq m$) be a rank- k incoherent matrix, i.e., both U^* and V^* are μ -incoherent (see Definition 2.4). Also, let each entry of M be observed uniformly and independently with probability,*

$$p > C \frac{\left(\frac{\sigma_1^*}{\sigma_k^*}\right)^2 \mu^2 k^{2.5} \log n \log \frac{k\|M\|_F}{\epsilon}}{m\delta_{2k}^2},$$

where $\delta_{2k} \leq \frac{\sigma_k^*}{12k\sigma_1^*}$ and $C > 0$ is a global constant. Then w.h.p. for $T = C' \log \frac{\|M\|_F}{\epsilon}$, the outputs \hat{U}^T and V^T of Algorithm 2, with input $(\Omega, P_\Omega(M))$ (see Equation (2)) satisfy: $\left\|M - \hat{U}^T (V^T)^\dagger\right\|_F \leq \epsilon$.

The above theorem implies that by observing $|\Omega| = O\left(\left(\frac{\sigma_1^*}{\sigma_k^*}\right)^4 k^{4.5} n \log n \log(k\|M\|_F/\epsilon)\right)$ random entries of an incoherent M , AltMinComplete can recover M in $O(\log(1/\epsilon))$ steps. In terms of sample complexity ($|\Omega|$), our results show alternating minimization may require a bigger Ω than convex optimization, as our result has $|\Omega|$ depend on the condition number, required accuracy (ϵ) and worse dependence on k than known bounds. In contrast, trace-norm minimization based methods require $O(kn \log n)$ samples only.

Empirically however, this is not seen to be the case [10] and we leave further tightening of the sample complexity bounds for matrix completion as an open problem.

In terms of time complexity, we show that AltMinComplete needs time $O(|\Omega|k^2 \log(1/\epsilon))$. This is in contrast to popular trace-norm minimization based methods that need $O(1/\sqrt{\epsilon})$ steps [1] and total time complexity of $O(|\Omega|n/\sqrt{\epsilon})$; note that the latter can be potentially quadratic in n . Furthermore, each step of such methods requires computation of the SVD of an $m \times n$ matrix. As mentioned earlier, interior point methods for trace-norm minimization also converge in $O(\log(1/\epsilon))$ steps but each iteration requires $O(n^5)$ steps and need storage of the entire $m \times n$ matrix X .

3 Related Work

Alternating Minimization: Alternating minimization and its variants have been applied to several low-rank matrix estimation problems. For example, clustering [14], sparse PCA [22], non-negative matrix factorization [13], signed network prediction [9] etc. There are three main reasons for such wide applicability of this approach: a) low-memory footprint and fast iterations, b) flexible modeling, c) amenable to parallelization. However, despite such empirical success, this approach has largely been used as a heuristic and has had no theoretical analysis other than the guarantees of convergence to the *local minima* [20]. Ours is the first analysis of this approach for two practically important problems: a) matrix completion, b) matrix sensing.

Matrix Completion: This is the problem of completing a low-rank matrix from a few sampled entries. Candes and Recht [3] provided the first results on this problem, showing that under the random sampling and incoherence conditions (detailed above), $O(kn^{1.2} \log n)$ samples allow for recovery via convex trace-norm minimization; this was improved to $O(kn \log n)$ in [5]. For large matrices, this approach is not very attractive due to the need to store and update the entire matrix, and because iterative methods for trace norm minimization require $O(\frac{1}{\sqrt{\epsilon}})$ steps to achieve additive error of ϵ . Moreover, each such step needs to compute an SVD.

Another approach, in [12], involved taking a single SVD, followed by gradient descent on a Grassmanian manifold. However, (a) this is more expensive than alternating minimization as it needs to compute gradient over Grassmanian manifold which in general is a computationally intensive step, and (b) the analysis of the algorithm only guarantees asymptotic convergence, and in the worst case might take exponential time in the problem size.

Recently, several other matrix completion type of problems have been studied in the literature. For example, robust PCA [6, 2], spectral clustering [11] etc. Here again, under additional assumptions, convex relaxation based methods have rigorous analysis but alternating minimization based algorithms continue to be algorithms of choice in practice.

Matrix Sensing: The general problem of matrix sensing was first proposed by [19]. They established recovery via trace norm minimization, assuming the sensing operator satisfies “restricted isometry” conditions. Subsequently, several other methods [10, 17] were proposed for this problem that also recovers the underlying matrix with optimal number of measurements and can give an ϵ -additive approximation in time $O(\log(1/\epsilon))$. But, similar to matrix completion, most of these methods require computing SVD of a large matrix at each step and hence have poor scalability to large problems.

We show that AltMinSense and AltMin-Completion provide more scalable algorithms for their respective problems. We demonstrate that these algorithms have geometric convergence to the optima, while each iteration is relatively cheap. For this, we assume conditions similar to those

required by existing algorithms; albeit, with one drawback: number of samples required by our analysis depend on the condition number of the underlying matrix M . For the matrix sensing problem, we remove this requirement by using a stagewise algorithm; we leave similar analysis for matrix completion as an open problem.

4 Matrix Sensing

In this section, we study the matrix sensing problem (LRMS) and prove that if the measurement operator, \mathcal{A} , satisfies RIP then AltMinSense (Algorithm 1) recovers the underlying low-rank matrix *exactly* (see Theorem 2.2).

At a high level, we prove Theorem 2.2 by showing that the “distance” between subspaces spanned by \widehat{V}^t (iterate at time t) and V^* decreases exponentially with t . This done based on the observation that once the (standard) matrix RIP condition (Definition 2.1) holds, alternating minimization can be viewed, and analyzed, as a **perturbed version of the power method**. This is easiest to see for the rank-1 case below; we detail this proof, and then the more general rank- k case.

In this paper, we use the following definition of distance between subspaces:

Definition 4.1. [8] *Given two matrices $\widehat{U}, \widehat{W} \in \mathbb{R}^{m \times k}$, the (principal angle) distance between the subspaces spanned by the columns of \widehat{U} and \widehat{W} is given by:*

$$\text{dist}(\widehat{U}, \widehat{W}) \stackrel{\text{def}}{=} \|U_{\perp}^{\dagger} W\|_2 = \|W_{\perp}^{\dagger} U\|_2$$

where U and W are orthonormal bases of the spaces $\text{Span}(\widehat{U})$ and $\text{Span}(\widehat{W})$, respectively. Similarly, U_{\perp} and W_{\perp} are any orthonormal bases of the perpendicular spaces $\text{Span}(U)^{\perp}$ and $\text{Span}(W)^{\perp}$, respectively.

Note: (a) The distance depends only on the spaces spanned by the columns of \widehat{U}, \widehat{W} , (b) if the ranks of \widehat{U} and \widehat{W} (i.e. the dimensions of their spans) are not equal, then $\text{dist}(\widehat{U}, \widehat{W}) = 1$, and (c) $\text{dist}(\widehat{U}, \widehat{W}) = 0$ if and only if they span the same subspace of \mathbb{R}^m .

We now present a theorem that bounds the distance between the subspaces spanned by \widehat{V}^t and V^* and show that it decreases exponentially with t .

Theorem 4.2. *Let $b = \mathcal{A}(M)$ where M and \mathcal{A} satisfy assumptions given in Theorem 2.2. Then, the $(t+1)$ -th iterates $\widehat{U}^{t+1}, \widehat{V}^{t+1}$ of AltMinSense satisfy:*

$$\begin{aligned} \text{dist}(\widehat{V}^{t+1}, V^*) &\leq \frac{1}{4} \cdot \text{dist}(\widehat{U}^t, U^*) , \\ \text{dist}(\widehat{U}^{t+1}, U^*) &\leq \frac{1}{4} \cdot \text{dist}(\widehat{V}^{t+1}, V^*) \end{aligned}$$

where $\text{dist}(U, W)$ denotes the principal angle based distance (see Definition 4.1).

Using Theorem 4.2, we are now ready to prove Theorem 2.2.

Proof Of Theorem 2.2. Assuming correctness of Theorem 4.2, Theorem 2.2 follows by using the following set of inequalities:

$$\begin{aligned}
\|M - \widehat{U}^T(\widehat{V}^T)^\dagger\|_F^2 &\stackrel{\zeta_1}{\leq} \frac{1}{1 - \delta_{2k}} \|\mathcal{A}(M - \widehat{U}^T(\widehat{V}^T)^\dagger)\|_2^2, \\
&\stackrel{\zeta_2}{\leq} \frac{1}{1 - \delta_{2k}} \|\mathcal{A}(M(I - V^T(V^T)^\dagger))\|_2^2, \\
&\stackrel{\zeta_3}{\leq} \frac{1 + \delta_{2k}}{1 - \delta_{2k}} \|U^* \Sigma^* (V^*)^\dagger (I - V^T(V^T)^\dagger)\|_F^2, \\
&\stackrel{\zeta_4}{\leq} \frac{1 + \delta_{2k}}{1 - \delta_{2k}} \|M\|_F^2 \text{dist}^2(V^T, V^*) \stackrel{\zeta_5}{\leq} \epsilon,
\end{aligned}$$

where V^T is an orthonormal basis of \widehat{V}^T , ζ_1 and ζ_3 follow by RIP, ζ_2 holds as \widehat{U}^T is the least squares solution, ζ_4 follows from the definition of $\text{dist}(\cdot, \cdot)$ and finally ζ_5 follows from Theorem 4.2 and by setting T appropriately. \square

To complete the proof of Theorem 2.2, we now need to prove Theorem 4.2. In the next section, we illustrate the main ideas of the proof of Theorem 4.2 by applying it to a rank-1 matrix i.e., when $k = 1$. We then provide a proof of Theorem 4.2 for arbitrary k in Section 4.2.

4.1 Rank-1 Case

In this section, we provide a proof of Theorem 4.2 for the special case of $k = 1$. That is, let $M = u^* \sigma^* (v^*)^\dagger$ s.t. $u^* \in \mathbb{R}^m$, $\|u^*\|_2 = 1$ and $v^* \in \mathbb{R}^n$, $\|v^*\|_2 = 1$. Also note that when \widehat{u} and \widehat{w} are vectors, $\text{dist}(\widehat{u}, \widehat{w}) = 1 - (u^\dagger w)^2$, where $u = \widehat{u}/\|\widehat{u}\|_2$ and $w = \widehat{w}/\|\widehat{w}\|_2$.

Consider the t -th update step in the AltMinSense procedure. As $\widehat{v}^{t+1} = \text{argmin}_{\widehat{v}} \sum_{i=1}^d \left(\widehat{u}^\dagger A_i^\dagger \widehat{v} - \sigma^* u^{*\dagger} A_i^\dagger v^* \right)^2$, setting the gradient of the above objective function to 0, we obtain:

$$\left(\sum_{i=1}^d A_i u^t (u^t)^\dagger A_i^\dagger \right) \|\widehat{u}^t\|_2 \widehat{v}^{t+1} = \sigma^* \left(\sum_{i=1}^d A_i u^t u^{*\dagger} A_i^\dagger \right) v^*,$$

where $u^t = \widehat{u}^t / \|\widehat{u}^t\|_2$. Now, let $B = \sum_{i=1}^d A_i u^t (u^t)^\dagger A_i^\dagger$ and $C = \sum_{i=1}^d A_i u^t (u^*)^\dagger A_i^\dagger$. Then,

$$\begin{aligned}
\|\widehat{u}^t\|_2 \widehat{v}^{t+1} &= \sigma^* B^{-1} C v^*, \\
&= \underbrace{\langle u^*, u^t \rangle \sigma^* v^*}_{\text{Power Method}} - \underbrace{B^{-1} (\langle u^*, u^t \rangle B - C) \sigma^* v^*}_{\text{Error Term}}.
\end{aligned} \tag{4}$$

Note that the first term in the above expression is the power method iterate (i.e., $M^\dagger u^t$). The second term is an error term and the goal is to show that it becomes smaller as u^t gets closer to u^* . Note that when $u^t = u^*$, the error term is 0 *irrespective* of the measurement operator \mathcal{A} .

Below, we provide a precise bound on the error term:

Lemma 4.3. *Consider the error term defined in (4) and let \mathcal{A} satisfy 2-RIP with constant δ_2 . Then,*

$$\|B^{-1} (\langle u^*, u^t \rangle B - C) v^*\| \leq \frac{3\delta_2}{1 - 3\delta_2} \sqrt{1 - \langle u^t, u^* \rangle^2}$$

See Appendix B.1 for a detailed proof of the above lemma.

Using the above lemma, we now finish the proof of Theorem 4.2:

Proof of Rank-1 case of Theorem 4.2. Let $v^{t+1} = \hat{v}^{t+1} / \|\hat{v}^{t+1}\|_2$. Now, using (4) and Lemma 4.3.,

$$\begin{aligned} \langle v^{t+1}, v^* \rangle &= \frac{\langle \hat{v}^{t+1}, v^* \rangle}{\|\hat{v}^{t+1}\|} = \frac{\langle \hat{v}^{t+1} / \sigma^*, v^* \rangle}{\|\hat{v}^{t+1} / \sigma^*\|} \\ &\leq \frac{\langle u^*, u^t \rangle - \hat{\delta}_2 \sqrt{1 - \langle u^*, u^t \rangle^2}}{\sqrt{\left(\langle u^*, u^t \rangle - \hat{\delta}_2 \sqrt{1 - \langle u^*, u^t \rangle^2} \right)^2 + \hat{\delta}_2^2 (1 - \langle u^*, u^t \rangle^2)}}, \end{aligned}$$

where $\hat{\delta}_2 = \frac{3\delta_2}{1-3\delta_2}$. That is,

$$\text{dist}^2(v^{t+1}, v^*) \leq \frac{\hat{\delta}_2^2 (1 - \langle u^*, u^t \rangle^2)}{(\langle u^*, u^t \rangle - \hat{\delta}_2 \sqrt{1 - \langle u^*, u^t \rangle^2})^2 + \hat{\delta}_2^2 (1 - \langle u^*, u^t \rangle^2)}.$$

Hence, assuming $\langle u^*, u^t \rangle \geq 5\hat{\delta}_2$, $\text{dist}(v^{t+1}, v^*) \leq \frac{1}{4} \text{dist}(u^t, u^*)$. As $\text{dist}(u^{t+1}, u^*)$ and $\text{dist}(v^{t+1}, v^*)$ are decreasing with t (from the above bound), we only need to show that $\langle u^0, u^t \rangle \geq 5\hat{\delta}_2$. Recall that \hat{u}^0 is obtained by using one step of SVP algorithm [10]. Hence, using Lemma 2.1 of [10] (see Lemma A.1):

$$\|\sigma_1^*(I - u^0(u^0)^\dagger)u^*\|_2^2 \leq \|M - \hat{u}^0(\hat{v}^0)^\dagger\|_F^2 \leq 2\delta_2 \|M\|_F^2.$$

Therefore, $\langle u^0, u^* \rangle \geq \sqrt{1 - 2\delta_2} \geq 5\hat{\delta}_2$ assuming $\delta_2 \leq \frac{1}{100}$. \square

4.2 Rank- k Case

In this section, we present the proof of Theorem 4.2 for arbitrary k , i.e., when M is a rank- k matrix (with SVD $U^* \Sigma^* (V^*)^\dagger$).

Similar to the analysis for the rank-1 case (Section 4.1), we show that even for arbitrary k , the updates of AltMinSense are essentially power-method type updates but with a bounded error term whose magnitude decreases with each iteration.

However, directly analyzing iterates of AltMinSense is a bit tedious due to non-orthonormality of intermediate iterates \hat{U} . Instead, **for analysis only** we consider the iterates of a modified version of AltMinSense, where we explicitly orthonormalize each iterate using the QR-decomposition². In particular, suppose we replace steps 4 and 5 of AltMinSense with the following

$$\begin{aligned} \hat{U}^t &= U^t R_U^t \quad (\text{QR decomposition}), \\ \hat{V}^{t+1} &\leftarrow \underset{V}{\text{argmin}} \|\mathcal{A}(U^t V^\dagger) - b\|_2^2, \\ \hat{V}^{t+1} &= V^{t+1} R_V^{t+1} \quad (\text{QR decomposition}) \\ \hat{U}^{t+1} &\leftarrow \underset{U}{\text{argmin}} \|\mathcal{A}(U(V^{t+1})^\dagger) - b\|_2^2 \end{aligned} \tag{5}$$

²The QR decomposition factorizes a matrix into an orthonormal matrix (a basis of its column space) and an upper triangular matrix; that is given \hat{S} it computes $\hat{S} = SR$ where S has orthonormal columns and R is upper triangular. If \hat{S} is full-rank, so are S and R .

In our algorithm, in each iterate both \hat{U}^t, \hat{V}^t remain full-rank because $\text{dist}(U^t, U^*) < 1$; with this, the following lemma implies that the spaces spanned by the iterates in our AltMinSense algorithm are *exactly the same* as the respective ones by the iterates of the above modified version (and hence the distances $\text{dist}(\hat{U}^t, U^*)$ and $\text{dist}(\hat{V}^t, V^*)$ are also the same for the two algorithms).

Lemma 4.4. *Let \hat{U}^t be the t^{th} iterate of our AltMinSense algorithm, and \tilde{U}^t of the modified version stated above. Suppose also that both \hat{U}^t, \tilde{U}^t are full-rank, and span the same subspace. Then the same will be true for the subsequent iterates for the two algorithms, i.e. $\text{Span}(\hat{V}^{t+1}) = \text{Span}(\tilde{V}^{t+1})$, $\text{Span}(\hat{U}^{t+1}) = \text{Span}(\tilde{U}^{t+1})$, and all matrices at iterate $t+1$ will be full-rank.*

The proof of the above lemma can be found in Appendix B.2. In light of this, we will now prove Theorem 4.2 with the new QR-based iterates (5).

Lemma 4.5. *Let \hat{U}^t be the t -th step iterate of AltMinSense and let U^t, \hat{V}^{t+1} and V^{t+1} be obtained by Update (5). Then,*

$$\hat{V}^{t+1} = \underbrace{V^* \Sigma^* U^{*\dagger} U^t}_{\text{Power-method Update}} - \underbrace{F}_{\text{Error Term}}, \quad V^{t+1} = \hat{V}^{t+1} (R^{(t+1)})^{-1}, \quad (6)$$

where F is an error matrix defined in (8) and $R^{(t+1)}$ is a triangular matrix obtained using QR-decomposition of \hat{V}^{t+1} .

See Appendix B for a detailed proof of the above lemma.

Before we give an expression for the error matrix F , we define the following notation. Let $v^* \in \mathbb{R}^{nk}$ be given by: $v^* = \text{vec}(V^*)$, i.e., $v^* = [v_1^{*\dagger} v_2^{*\dagger} \dots v_k^{*\dagger}]^\dagger$. Define B, C, D, S as follows:

$$B \stackrel{\text{def}}{=} \begin{bmatrix} B_{11} & \dots & B_{1k} \\ \vdots & \ddots & \vdots \\ B_{k1} & \dots & B_{kk} \end{bmatrix}, \quad C \stackrel{\text{def}}{=} \begin{bmatrix} C_{11} & \dots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{k1} & \dots & C_{kk} \end{bmatrix},$$

$$D \stackrel{\text{def}}{=} \begin{bmatrix} D_{11} & \dots & D_{1k} \\ \vdots & \ddots & \vdots \\ D_{k1} & \dots & D_{kk} \end{bmatrix}, \quad S \stackrel{\text{def}}{=} \begin{bmatrix} \sigma_1^* I_n & \dots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \dots & \sigma_k^* I_n \end{bmatrix}. \quad (7)$$

where, for $1 \leq p, q \leq k$: $B_{pq} \stackrel{\text{def}}{=} \sum_{i=1}^d A_i u_p^t u_q^{t\dagger} A_i^\dagger$, $C_{pq} \stackrel{\text{def}}{=} \sum_{i=1}^d A_i u_p^t u_q^{*\dagger} A_i^\dagger$, and, $D_{pq} \stackrel{\text{def}}{=} \langle u_p^t, u_q^* \rangle \mathbb{I}_{n \times n}$. Recall that, u_p^t is the p -th column of U^t and u_q^* is the q -th left singular vector of the underlying matrix $M = U^* \Sigma^* (V^*)^\dagger$. Finally F is obtained by “de-stacking” the vector $B^{-1} (BD - C) S v^*$ i.e., the i^{th} column of F is given by:

$$F_i \stackrel{\text{def}}{=} \begin{bmatrix} (B^{-1} (BD - C) S v^*)_{ni+1} \\ (B^{-1} (BD - C) S v^*)_{ni+2} \\ \vdots \\ (B^{-1} (BD - C) S v^*)_{ni+n} \end{bmatrix}, \quad F \stackrel{\text{def}}{=} [F_1 \ F_2 \ \dots \ F_k]. \quad (8)$$

Note that the notation above should have been B^t, C^t and so on. We suppress the dependence on t for notational simplicity. Now, from Update (6), we have

$$V^{t+1} = \hat{V}^{t+1} R^{(t+1)-1} = (V^* \Sigma^* U^{*\dagger} U^t - F) R^{(t+1)-1}$$

$$\Rightarrow V_\perp^{*\dagger} V^{t+1} = -V_\perp^{*\dagger} F R^{(t+1)-1}. \quad (9)$$

where V_\perp^* is an orthonormal basis of $\text{Span}(v_1^*, v_2^*, \dots, v_k^*)^\perp$. Therefore,

$$\text{dist}(V^*, V^{t+1}) = \|V_\perp^{*\dagger} V^{t+1}\|_2 = \|V_\perp^{*\dagger} F R^{(t+1)^{-1}}\|_2 \leq \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^* R^{(t+1)^{-1}}\|_2.$$

Now, we break down the proof of Theorem 4.2 into the following two steps:

- show that $\|F(\Sigma^*)^{-1}\|_2$ is small (Lemma 4.6) and
- show that $\|\Sigma^* R^{(t+1)^{-1}}\|_2$ is small (Lemma 4.7).

We will now state the two corresponding lemmas. Complete proofs can be found in Appendix B.2 The first lemma bounds the spectral norm of $F(\Sigma^*)^{-1}$.

Lemma 4.6. *Let linear measurement \mathcal{A} satisfy RIP for all $2k$ -rank matrices and let $b = \mathcal{A}(M)$ with $M \in \mathbb{R}^{m \times n}$ being a rank- k matrix. Then, spectral norm of error matrix $F(\Sigma^*)^{-1}$ (see Equation 6) after t -th iteration update satisfy:*

$$\|F(\Sigma^*)^{-1}\|_2 \leq \frac{\delta_{2k} k}{1 - \delta_{2k}} \text{dist}(U^t, U^*). \quad (10)$$

The following lemma bounds the spectral norm of $\Sigma^* R^{(t+1)^{-1}}$.

Lemma 4.7. *Let linear measurement \mathcal{A} satisfy RIP for all $2k$ -rank matrices and let $b = \mathcal{A}(M)$ with $M \in \mathbb{R}^{m \times n}$ being a rank- k matrix. Then,*

$$\|\Sigma^* (R^{(t+1)})^{-1}\|_2 \leq \frac{\sigma_1^*/\sigma_k^*}{\sqrt{1 - \text{dist}^2(U^t, U^*) - \frac{(\sigma_1^*/\sigma_k^*)\delta_{2k}k\text{dist}(U^t, U^*)}{1 - \delta_{2k}}}}. \quad (11)$$

With the above two lemmas, we now prove Theorem 4.2.

Proof Of Theorem 4.2. Using (9), (10) and (11), we obtain the following:

$$\begin{aligned} \text{dist}(V^{t+1}, V^*) &= \|V_\perp^{*\dagger} V^{t+1}\|_2, \\ &\leq \|V_\perp^{*\dagger} F(\Sigma^*)^{-1} \Sigma^* R^{(t+1)^{-1}}\|_2, \\ &\leq \|V_\perp^*\|_2 \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^* R^{(t+1)^{-1}}\|_2 \\ &\leq \frac{(\sigma_1^*/\sigma_k^*)\delta_{2k}k \cdot \text{dist}(U^t, U^*)}{(1 - \delta_{2k})L}, \end{aligned} \quad (12)$$

where $L = \sqrt{1 - \text{dist}(U^t, U^*)^2 - \frac{(\sigma_1^*/\sigma_k^*)\delta_{2k}k\text{dist}(U^t, U^*)}{1 - \delta_{2k}}}$. Also, note that U^0 is obtained using SVD of $\sum_i A_i b_i$. Hence, using Lemma A.1, we have:

$$\begin{aligned} \|\mathcal{A}(U^0 \Sigma^0 V^0 - U^* \Sigma^* (V^*)^\dagger)\|_2^2 &\leq 4\delta_{2k} \|\mathcal{A}(U^* \Sigma^* (V^*)^\dagger)\|_2^2, \\ \Rightarrow \|U^0 \Sigma^0 V^0 - U^* \Sigma^* (V^*)^\dagger\|_F^2 &\leq 4\delta_{2k} (1 + 3\delta_{2k}) \|\Sigma^*\|_F^2, \\ \Rightarrow \|U^0 (U^0)^\dagger U^* \Sigma^* (V^*)^\dagger - U^* \Sigma^* (V^*)^\dagger\|_F^2 &\leq 6\delta_{2k} \|\Sigma^*\|_F^2, \\ \Rightarrow (\sigma_k^*)^2 \|(U^0 (U^0)^\dagger - I) U^*\|_F^2 &\leq 6\delta_{2k} k (\sigma_1^*)^2, \\ \Rightarrow \text{dist}(U^0, U^*) &\leq \sqrt{6\delta_{2k} k} \left(\frac{\sigma_1^*}{\sigma_k^*} \right) < \frac{1}{2}. \end{aligned} \quad (13)$$

Using (12) with $\text{dist}(U^0, U^*) < \frac{1}{2}$ and $\delta_{2k} < \frac{1}{24(\sigma_1^*/\sigma_k^*)^2 k}$, we obtain: $\text{dist}(V^t, V^*) < \frac{1}{4} \text{dist}(U^t, U^*)$. Similarly we can show that $\text{dist}(U^{t+1}, U^*) < \frac{1}{4} \text{dist}(V^t, V^*)$. \square

5 Matrix Completion

In this section, we study the Matrix Completion problem (LRMC) and show that, assuming k and $\frac{\sigma_1^*}{\sigma_k^*}$ are constant, AltMinComplete (Algorithm 2) recovers the underlying matrix M using only $O(n \log n)$ measurements (i.e., we prove Theorem 2.5).

As mentioned, while observing elements in Ω constitutes a linear map, matrix completion is different from matrix sensing because the map does not satisfy RIP. The (now standard) approach is to assume incoherence of the true matrix M , as done in Definition 2.4. With this, and the random sampling of Ω , matrix completion exhibits similarities to matrix sensing. For our analysis, we can again use the fact that incoherence allows us to view alternating minimization as a perturbed power method, whose error we can control.

However, there are important differences between the two problems, which make the analysis of completion more complicated. Chief among them is the fact that we need to establish *the incoherence of each iterate*. For the first initialization \hat{U}^0 , this necessitates the “clipping” procedure (described in step 4 of the algorithm). For the subsequent steps, this requires the partitioning of the observed Ω into $2T + 1$ sets (as described in step 2 of the algorithm).

As in the case of matrix sensing, we prove our main result for matrix completion (Theorem 2.5) by first establishing a geometric decay of the distance between the subspaces spanned by \hat{U}^t, \hat{V}^t and U^*, V^* respectively.

Theorem 5.1. *Under the assumptions of Theorem 2.5, the $(t+1)^{th}$ iterates \hat{U}^{t+1} and \hat{V}^{t+1} satisfy the following property w.h.p.:*

$$\begin{aligned} \text{dist}(\hat{V}^{t+1}, V^*) &\leq \frac{1}{4} \text{dist}(\hat{U}^t, U^*) \quad \text{and} \\ \text{dist}(\hat{U}^{t+1}, U^*) &\leq \frac{1}{4} \text{dist}(\hat{V}^{t+1}, V^*), \quad \forall 1 \leq t \leq T. \end{aligned}$$

We use the above result along with incoherence of M to prove Theorem 2.5. See Appendix C for a detailed proof.

Now, similar to the matrix sensing case, alternating minimization needs an initial iterate that is close enough to U^* and V^* , from where it will then converge. To this end, Steps 3–4 of Algorithm 2 use SVD of $P_\Omega(M)$ followed by clipping to initialize \hat{U}^0 . While the SVD step guarantees that \hat{U}^0 is close enough to U^* , it might not remain incoherent. To maintain incoherence, we introduce an extra clipping step which guarantees incoherence of \hat{U}^0 while also ensuring that \hat{U}^0 is close enough to U^* (see Lemma 5.2)

Lemma 5.2. *Let M, Ω, p be as defined in Theorem 2.5. Also, let U^0 be the initial iterate obtained by step 4 of Algorithm 2. Then, w.h.p. we have*

- $\text{dist}(U^0, U^*) \leq \frac{1}{2}$ and
- U^0 is incoherent with parameter $4\mu\sqrt{k}$.

The above lemma guarantees a “good” starting point for alternating minimization. Using this, we now present a proof of Theorem 5.1. Similar to the sensing section, we first explain key ideas of our proof using rank-1 example. Then in Section 5.2 we extend our proof to general rank- k matrices.

5.1 Rank-1 Case

Consider the rank-1 matrix completion problem where $M = \sigma^* u^* (v^*)^\dagger$. Now, the t -th step iterates \hat{v}^{t+1} of Algorithm 2 are given by:

$$\hat{v}^{t+1} = \underset{\hat{v}}{\operatorname{argmin}} \sum_{(i,j) \in \Omega} (M_{ij} - \hat{u}_i^t \hat{v}_j)^2.$$

Let $u^t = \hat{u}^t / \|\hat{u}^t\|_2$. Then, $\forall j$:

$$\begin{aligned} \|\hat{u}^t\|_2 \sum_{i:(i,j) \in \Omega} (u_i^t)^2 \hat{v}_j^{t+1} &= \sigma^* \sum_{i:(i,j) \in \Omega} u_i^t u_i^* v_j^* \\ \Rightarrow \|\hat{u}^t\|_2 \hat{v}_j^{t+1} &= \frac{\sigma^*}{\sum_{i:(i,j) \in \Omega} (u_i^t)^2} \sum_{i:(i,j) \in \Omega} u_i^t u_i^* v_j^* \\ &= \sigma^* \langle u^t, u^* \rangle v_j^* - \frac{\sigma^* (\langle u^t, u^* \rangle \sum_{i:(i,j) \in \Omega} (u_i^t)^2 v_j^* - \sum_{i:(i,j) \in \Omega} u_i^t u_i^* v_j^*)}{\sum_{i:(i,j) \in \Omega} (u_i^t)^2}. \end{aligned} \quad (14)$$

Hence,

$$\|\hat{u}^t\|_2 \hat{v}^{t+1} = \underbrace{\langle u^*, u^t \rangle \sigma^* v^*}_{\text{Power Method}} - \underbrace{\sigma^* B^{-1} (\langle u^t, u^* \rangle B - C) v^*}_{\text{Error Term}}, \quad (15)$$

where $B, C \in \mathbb{R}^{n \times n}$ are diagonal matrices, such that,

$$B_{jj} = \frac{\sum_{i:(i,j) \in \Omega} (u_i^t)^2}{p}, \quad C_{jj} = \frac{\sum_{i:(i,j) \in \Omega} u_i^t u_i^*}{p}. \quad (16)$$

Note the similarities between the update (15) and the rank-1 update (4) for the sensing case. Here again, it is essentially a power-method update (first term) along with a bounded error term (see Lemma 5.3). Using this insight, we now prove Theorem 5.1 for the special case of rank-1 matrices. Our proof can be divided in three major steps:

- *Base Case*: Show that $u^0 = \hat{u}^0 / \|\hat{u}^0\|_2$ is incoherent and have small distance to u^* (see Lemma 5.2).
- *Induction Step (distance)*: Assuming $u^t = \hat{u}^t / \|\hat{u}^t\|_2$ to be incoherent and that u^t has a small distance to u^* , v^{t+1} decreases distances to v^* by at least a constant factor.
- *Induction Step (incoherence)*: Show incoherence of v^{t+1} , while assuming incoherence of u^t (see Lemma 5.4)

We first prove the second step of our proof. To this end, we provide the following lemma that bounds the error term. See Appendix C.2 for a proof of the below given lemma.

Lemma 5.3. *Let M, p, Ω, u^t be as defined in Theorem 2.5. Also, let u^t be a unit vector with incoherence parameter $\mu_1 = \frac{6(1+\delta_2)\mu}{1-\delta_2}$. Then, w.p. at least $1 - \frac{1}{n^3}$:*

$$\|B^{-1} (\langle u^*, u^t \rangle B - C) v^*\|_2 \leq \frac{\delta_2}{1 - \delta_2} \sqrt{1 - \langle u^t, u^* \rangle^2}.$$

Multiplying (15) with v^* and using Lemma 5.3, we get:

$$\|\hat{u}^t\|_2 \langle \hat{v}^{t+1}, v^* \rangle \geq \sigma^* \langle u^t, u^* \rangle - 2\sigma^* \delta_2 \sqrt{1 - \langle u^t, u^* \rangle^2}, \quad (17)$$

where $\delta_2 < \frac{1}{12}$ is a constant defined in the Theorem statement and is similar to the RIP constant in Section 4.

Similarly, by multiplying (15) with v_\perp (where $\langle v_\perp^*, v^* \rangle = 0$ and $\|v_\perp^*\|_2 = 1$) and using Lemma 5.3:

$$\|\hat{u}^t\|_2 \langle \hat{v}^{t+1}, v_\perp^* \rangle \leq 2\sigma^* \delta_2 \sqrt{1 - \langle u^t, u^* \rangle^2}.$$

Using the above two equations:

$$1 - \langle v^{t+1}, v^* \rangle^2 \leq \frac{4\delta_2^2(1 - \langle u^t, u^* \rangle^2)}{(\langle u^t, u^* \rangle - 2\delta_2 \sqrt{1 - \langle u^t, u^* \rangle^2})^2 + (2\delta_2 \sqrt{1 - \langle u^t, u^* \rangle^2})^2}.$$

Assuming, $\langle v^{t+1}, v^* \rangle \geq 6\delta_2$,

$$\text{dist}(v^{t+1}, v^*) = \sqrt{1 - \langle v^{t+1}, v^* \rangle^2} \leq \frac{1}{4} \sqrt{1 - \langle u^t, u^* \rangle^2}.$$

Using same arguments, we can show that, $\text{dist}(u^{t+1}, u^*) \leq \text{dist}(v^{t+1}, v^*)/4$. Hence, after $O(\log(1/\epsilon))$ iterations, $\text{dist}(u^t, u^*) \leq \epsilon$ and $\text{dist}(v^{t+1}, v^*) \leq \epsilon$. This proves our second step.

We now provide the following lemma to prove the third step. We stress that v^{t+1} does not increase the incoherence parameter (μ_1) when compared to that of u^t .

Lemma 5.4. *Let M , p , Ω be as defined in Theorem 2.5. Also, let u^t be a unit vector with incoherence parameter $\mu_1 = \frac{6(1+\delta_2)\mu}{1-\delta_2}$. Then, w.p. at least $1 - \frac{1}{n^3}$, v^{t+1} is also μ_1 incoherent.*

See Appendix C.2 for a detailed proof of the lemma.

Finally, for the base case we need that u^0 is μ_1 incoherent and also $\langle u^0, u^* \rangle \geq 6\delta_2$. This follows directly by using Lemma 5.2 and the fact that $\delta_2 \leq 1/12$.

Note that, to obtain an error of ϵ , AltMinComplete needs to run for $O\left(\log \frac{\|M\|_F}{\epsilon}\right)$ iterations. Also, we need to sample a fresh Ω at each iteration of AltMinComplete. Hence, the total number of samples needed by AltMinComplete is $O\left(\log \frac{\|M\|_F}{\epsilon}\right)$ larger than the number of samples required per step.

5.2 Rank- k case

We now extend our proof of Theorem 5.1 to matrices with arbitrary rank. Here again, we show that the AltMinComplete algorithm reduces to power method with bounded perturbation at each step.

Similar to the matrix sensing case, we analyze the following QR decomposition based update instead of directly analyzing the updates of Algorithm 2:

$$\begin{aligned} \hat{U}^t &= U^t R_U^t \quad (\text{QR decomposition}), \\ \hat{V}^{t+1} &= \underset{\hat{V}}{\text{argmin}} \|P_\Omega(U^t \hat{V}^\dagger) - P_\Omega(M)\|_F^2, \\ \hat{V}^{t+1} &= V^{t+1} R_V^{t+1}. \quad (\text{QR decomposition}), \\ \hat{U}^{t+1} &= \underset{\hat{U}}{\text{argmin}} \|P_\Omega(\hat{U}(V^{t+1})^\dagger) - P_\Omega(M)\|_F^2. \end{aligned} \quad (18)$$

Here again, we would stress that the updates output exactly the same matrices at the end of each iteration and we prefer QR-based updates due to notational ease.

Now, as matrix completion is a special case of matrix sensing, Lemma 4.5 characterizes the updates of the AltMinComplete algorithm (see Algorithm 2). That is,

$$\begin{aligned}\widehat{V}^{t+1} &= \underbrace{V^* \Sigma^* U^{*\dagger} U^t}_{\text{Power-method Update}} - \underbrace{F}_{\text{Error Term}}, \\ V^{t+1} &= \widehat{V}^{t+1} (R^{(t+1)})^{-1},\end{aligned}\tag{19}$$

where F is the error matrix defined in (8) and $R^{(t+1)}$ is a upper-triangular matrix obtained using QR-decomposition of \widehat{V}^{t+1} . See (7) for the definition of B, C, D , and S .

Also, note that for the special case of matrix completion, $B_{pq}, C_{pq}, 1 \leq p, q \leq k$ are *diagonal matrices* with

$$(B_{pq})_{jj} = \frac{1}{p} \sum_{i:(i,j) \in \Omega} U_{ip}^t U_{iq}^t, \quad (C_{pq})_{jj} = \frac{1}{p} \sum_{i:(i,j) \in \Omega} U_{ip}^t U_{iq}^*.$$

We use this structure to further simplify the update equation. We first define matrices $B^j, C^j, D^j \in \mathbb{R}^{k \times k}, 1 \leq j \leq n$:

$$B^j = \frac{1}{p} \sum_{i:(i,j) \in \Omega} (U^t)^{(i)} (U^t)^{(i)\dagger}, \quad C^j = \frac{1}{p} \sum_{i:(i,j) \in \Omega} (U^t)^{(i)} (U^*)^{(i)\dagger},$$

and $D^j = (U^t)^\dagger U^*$. Using the above notation, (19) decouples into n equations of the form ($1 \leq j \leq n$):

$$(V^{t+1})^{(j)} = (V^*)^{(j)} (D^j - (B^j)^{-1} (B^j D^j - C^j)) (R^{(t+1)})^{-1},\tag{20}$$

where $(V^{t+1})^{(j)}$ and $(V^*)^{(j)}$ denote the j^{th} rows of V^{t+1} and V^* respectively.

Using the above notation, we now provide a proof of Theorem 5.1 for the general rank- k case.

Proof of Theorem 5.1. Multiplying the update equation (19) on the left by $(V_\perp^*)^\dagger$, we get: $(V_\perp^*)^\dagger \widehat{V}^{t+1} = -(V_\perp^*)^\dagger F (R^{(t+1)})^{-1}$. That is,

$$\begin{aligned}\text{dist}(V^*, V^{t+1}) &= \|V_\perp^{*\dagger} V^{(t+1)}\|_2 = \|V_\perp^{*\dagger} F R^{(t+1)-1}\|_2 \\ &\leq \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^* R^{(t+1)-1}\|_2.\end{aligned}$$

Now, similar to the sensing case (see Section 4.2) we break down our proof into the following two steps:

- Bound $\|F(\Sigma^*)^{-1}\|_2$ (Lemma 5.6) and
- Bound $\|\Sigma^* R^{(t+1)-1}\|_2$, i.e., the minimum singular value of $(\Sigma^*)^{-1} R^{(t+1)}$ (Lemma 5.7).

Using Lemma 5.6 and Lemma 5.7, w.p. at least $1 - 1/n^3$,

$$\text{dist}(V^*, V^{t+1}) \leq \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^* R^{(t+1)-1}\|_2 \leq \frac{(\sigma_1^*/\sigma_k^*)k(\delta_{2k}/(1-\delta_{2k})) \cdot \text{dist}(U^{(t)}, U^*)}{\sqrt{1 - \text{dist}(U^{(t)}, U^*)^2 - \frac{(\sigma_1^*/\sigma_k^*)k\delta_{2k}\text{dist}(U^{(t)}, U^*)}{1-\delta_{2k}}}}.$$

Now, using Lemma 5.2 we get: $\text{dist}(U^t, U^*) \leq \text{dist}(U^0, U^*) \leq \frac{1}{2}$. By selecting $\delta_{2k} < \frac{\sigma_k^*}{12k\sigma_1^*}$, i.e., $p \geq \frac{C(\sigma_1^*)^2 k^4 \log n}{m(\sigma_k^*)^2}$ and using above two inequalities:

$$\text{dist}(V^{t+1}, V^*) \leq \frac{1}{4} \text{dist}(U^t, U^*).$$

Furthermore, using Lemma 5.5 we get that V^{t+1} is μ_1 incoherent. Hence, using similar arguments as above, we also get: $\text{dist}(U^{t+1}, U^*) \leq (\frac{1}{4}) \text{dist}(V^{t+1}, V^*)$. \square

We now provide lemmas required by our above given proof. See Appendix C.3 for detailed proof of each of the lemmas.

We first provide a lemma to bound incoherence of V^{t+1} , assuming incoherence of U^t .

Lemma 5.5. *Let M, Ω, p be as defined in Theorem 2.5. Also, let U^t be the t -th step iterate obtained by (18). Let U^t be $\mu_1 = \frac{16\sigma_1^* \mu \sqrt{k}}{\sigma_k^*}$ incoherent. Then, w.p. at least $1 - 1/n^3$, iterate $V^{(t+1)}$ is also μ_1 incoherent.*

We now bound the error term (F) in AltMin update (19).

Lemma 5.6. *Let F be the error matrix defined by (8) (also see (19)) and let U^t be a μ_1 -incoherent orthonormal matrix obtained after $(t-1)^{\text{th}}$ update. Also, let M, Ω , and p satisfy assumptions of Theorem 2.5. Then, w.p. at least $1 - 1/n^3$:*

$$\|F(\Sigma^*)^{-1}\|_2 \leq \frac{\delta_{2k} k}{1 - \delta_{2k}} \text{dist}(U^t, U^*).$$

Next, we present a lemma to bound $\|(R^{(t+1)})^{-1}\|_2$.

Lemma 5.7. *Let $R^{(t+1)}$ be the lower-triangular matrix obtained by QR decomposition of \widehat{V}^{t+1} (see (19)) and let U^t be a μ_1 -incoherent orthonormal matrix obtained after $(t-1)^{\text{th}}$ update. Also, let M and Ω satisfy assumptions of Theorem 2.5. Then,*

$$\|\Sigma^*(R^{(t+1)})^{-1}\|_2 \leq \frac{\sigma_1^*/\sigma_k^*}{\sqrt{1 - \text{dist}^2(U^t, U^*) - \frac{(\sigma_1^*/\sigma_k^*)\delta_{2k} k \text{dist}(U^t, U^*)}{1 - \delta_{2k}}}}} \quad (21)$$

Proof. Lemma follows by exactly the same proof as that of Lemma 4.7 for the matrix sensing case. \square

6 Stagewise AltMin Algorithm

In Section 4, we showed that if $\delta_{2k} \leq \frac{(\sigma_k^*)^2}{(\sigma_1^*)^2 k}$ then AltMinSense (Algorithm 1) recovers the underlying matrix. This means that, $d = \frac{(\sigma_1^*)^4}{(\sigma_k^*)^4} k^2 n \log n$ random Gaussian measurements (assume $m \leq n$) are required to recover M . For matrices with large condition number (σ_1^*/σ_k^*), this would be significantly larger than the information theoretic bound of $O(kn \log n/k)$ measurements.

To alleviate this problem, we present a modified version of AltMinSense called Stage-AltMin. Stage-AltMin proceeds in k stages where in the i -th stage, a rank- i problem is solved. The goal of the i -th stage is to recover top i -singular vectors of M , up to $O(\sigma_{i+1}^*)$ error.

Algorithm 3 Stage-AltMin: Stagewise Alternating Minimization for Matrix Sensing

```

1: Input:  $b, \mathcal{A}$ 
2:  $\hat{U}^T \leftarrow \emptyset, \hat{V}^T \leftarrow \emptyset$ 
3: for  $i = 1, \dots, k$  do
4:    $[\hat{U}_{1:i}^0, \hat{V}_{1:i}^0] = \text{top } i\text{-singular vectors of } \left( \hat{U}_{1:i-1}^T (\hat{V}_{1:i-1}^T)^\dagger - \frac{3}{4} \mathcal{A}^T(\mathcal{A}(\hat{U}_{1:i-1}^T (\hat{V}_{1:i-1}^T)^\dagger) - b) \right)$  i.e.,
      one step of SVP [10]
5:   for  $t = 0, \dots, T-1$  do
6:      $\hat{V}_{1:i}^{t+1} \leftarrow \operatorname{argmin}_{V \in \mathbb{R}^{n \times i}} \|\mathcal{A}(\hat{U}_{1:i}^t V^\dagger) - b\|_2^2$ 
7:      $\hat{U}_{1:i}^{t+1} \leftarrow \operatorname{argmin}_{U \in \mathbb{R}^{m \times i}} \|\mathcal{A}(U_{1:i} (\hat{V}_{1:i}^{t+1})^\dagger) - b\|_2^2$ 
8:   end for
9: end for
10: Output:  $X = \hat{U}_{1:i}^T (\hat{V}_{1:i}^T)^\dagger$ 

```

Specifically, we initialize the i -th stage of our algorithm using one step of the SVP algorithm [10] (see Step 4 of Algorithm 3). We then show that, if $\delta_{2k} \leq \frac{1}{10k}$, then Stage-AltMin (Steps 6, 7 of Algorithm 3) decreases the error $\|M - \hat{U}_{1:i}^T (\hat{V}_{1:i}^T)^\dagger\|_F$ to $O(\sigma_{i+1}^*)$. Hence, after k steps, the error decreases to $O(\sigma_{k+1}^*) = 0$. Note that, $\hat{U}_{1:i}^t \in \mathbb{R}^{m \times i}$ represents the t -th step iterate (U) in the i -th stage; $\hat{V}_{1:i}^t \in \mathbb{R}^{n \times i}$ is also defined similarly.

Recall that, the main problem with our analysis of AltMinSense is that if $\sigma_i \gg \sigma_{i+1}$ (for some i) then $\delta_{2k} \leq \frac{(\sigma_{i+1}^*)^2}{(\sigma_i^*)^2 k}$ would need to be small. However, in such a scenario, the i -th stage of Algorithm 3 can be thought of as solving a noisy sensing problem where the goal is to recover $M_i \stackrel{\text{def}}{=} U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger$ using noisy measurements $b = \mathcal{A}(U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger) + N$ where noise matrix $N \stackrel{\text{def}}{=} U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger$. Here M_i and N represent the top i singular components and last $k-i$ singular components of M respectively. Hence, using noisy-case type analysis (see Section B.3) we show that the error $\|M - \hat{U}^t (\hat{V}^t)^\dagger\|_F$ decreases to $O(\sigma_{i+1}^*)$.

We now formally present the proof of our main result (see Theorem 2.3).

Proof Of Theorem 2.3. We prove the theorem using mathematical induction.

Base Case: After the 0-th step, error is: $\|M\|_F^2 \leq \sum_{j=1}^k \sigma_j^2 \leq k\sigma_1^2$. Hence, base case holds.

Induction Step: Here, assuming that the error bound holds for $(i-1)$ -th stage, we prove the error bound for the i -th stage.

Our proof proceeds in two steps. First, we show that the initial point $\hat{U}_{1:i}^0, \hat{V}_{1:i}^0$ of the i -th stage, obtained using Step 4, has $c(\sigma_i^*)^2 + O(k(\sigma_{i+1}^*)^2)$ error, with $c < 1$. In the second step, we show that using the initial points $\hat{U}_{1:i}^0, \hat{V}_{1:i}^0$, the AltMin algorithm iterations in the i -th stage (Steps 6, 7) reduces the error to $\max(\epsilon, 16k\sigma_{i+1}^2)$.

We formalize the above mentioned first step in Lemma 6.1 and then prove the second step in Lemma 6.2. \square

We now present two lemmas used by the above given proof. See Appendix B.4 for a proof of each of the lemmas.

Lemma 6.1. *Let assumptions of Theorem 2.3 be satisfied. Also, let $\hat{U}_{1:i}^0, \hat{V}_{1:i}^0$ be the output of*

Step 4 of Algorithm 3. Then, assuming that $\|M - \widehat{U}_{1:i-1}^T \widehat{V}_{1:i-1}^T\|_F^2 \leq 16k(\sigma_i^*)^2$, we obtain:

$$\left\|M - \widehat{U}_{1:i}^0 (\widehat{V}_{1:i}^0)^\dagger\right\|_F^2 \leq \sum_{j=i+1}^k (\sigma_j^*)^2 + \frac{1}{100}(\sigma_i^*)^2.$$

Lemma 6.2. Let assumptions of Theorem 2.3 be satisfied. Also, let $\widehat{U}_{1:i}^T, \widehat{V}_{1:i}^T$ be the T -th step iterates of the i -th stage of Algorithm 3. Then, assuming that $\|M - \widehat{U}_{1:i}^0 \widehat{V}_{1:i}^0\|_F^2 \leq \sum_{j=i+1}^k (\sigma_j^*)^2 + \frac{1}{100}(\sigma_i^*)^2$, we obtain:

$$\left\|M - \widehat{U}_{1:i}^T (\widehat{V}_{1:i}^T)^\dagger\right\|_F^2 \leq \max(\epsilon, 16k(\sigma_{i+1}^*)^2),$$

where $T = \Omega(\log(\|M\|_F/\epsilon))$.

7 Summary and Discussion

Alternating minimization provides an empirically appealing and popular approach to solving several different low-rank matrix recovery problems. The main motivation, and result, of this paper was to provide the **first theoretical guarantees on the global optimality of alternating minimization**, for matrix completion and the related problem of matrix sensing. We would like to note the following aspects of our results and proofs:

- For both the problems, we show that alternating minimization recovers the true matrix under *similar problem conditions* (RIP, incoherence) to those used by existing algorithms (based on convex optimization or iterated SVDs); computationally, our results show faster convergence to the global optima, but with possibly higher statistical (i.e. sample) complexity.
- We develop a new framework for analyzing alternating minimization for low-rank problems. Key observation of our framework is that for some problems (under standard problem conditions) alternating minimization can be viewed as a perturbed version of the power method. In our case, we can control the perturbation error based on the extent of RIP / incoherence demonstrated by the problem. This idea is likely to have applications to other similar problems where trace-norm based convex relaxation techniques have rigorous theoretical results but alternating minimization has enjoyed more empirical success. For example, robust PCA [6, 2], spectral clustering [11] etc.
- Our analysis also sheds light on two key aspects of the alternating minimization approach:
Initialization: Due to its connection to power method, it is now easy to see that for alternating minimization to succeed, the initial iterate should not be orthogonal to the target vector. Our results indeed show that alternating minimization succeeds if the initial iterate is not “almost orthogonal” to the target subspace. This suggests that, selecting initial iterate smartly is preferable to random initialization.

Dependence on the condition number: Our results for the alternating minimization algorithm depend on the condition number. However, using a stagewise adaptation of alternating minimization, we can remove this dependence for the matrix sensing problem. This suggests that (problem specific) modifications of the basic alternating minimization algorithm may in fact perform better than the original one, while (mostly) retaining the computational / implementational simplicity of the underlying method.

References

- [1] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.
- [2] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.
- [3] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, December 2009.
- [4] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
- [5] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.
- [6] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.
- [7] C. Chen, B. He, and X. Yuan. Matrix completion via an alternating direction method. *IMA Journal of Numerical Analysis*, 32(1):227–245, 2012.
- [8] G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.
- [9] C.-J. Hsieh, K.-Y. Chiang, and I. S. Dhillon. Low rank modeling of signed networks. In *KDD*, pages 507–515, 2012.
- [10] P. Jain, R. Meka, and I. S. Dhillon. Guaranteed rank minimization via singular value projection. In *NIPS*, pages 937–945, 2010.
- [11] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. In *ICML*, pages 1001–1008, 2011.
- [12] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.
- [13] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix Anal. Appl.*, 30(2):713–730, July 2008.
- [14] J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. Technical Report GT-CSE-08-01, Georgia Institute of Technology, 2008.
- [15] Y. Koren. The BellKor solution to the Netflix grand prize, 2009.
- [16] Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.
- [17] K. Lee and Y. Bresler. Admira: atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

- [18] R. Meka, P. Jain, C. Caramanis, and I. S. Dhillon. Rank minimization via online learning. In *ICML*, pages 656–663, 2008.
- [19] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [20] W. I. Zangwill. *Nonlinear Programming: A Unified Approach*. Englewood Cliffs: Prentice-Hall, 1969.
- [21] Y. Zhou, D. M. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *AAIM*, pages 337–348, 2008.
- [22] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *JCGS*, 15(2):262–286, 2006.

A Preliminaries

Lemma A.1 (Lemma 2.1 of [10]). *Let $b = \mathcal{A}(M) + e$, where e is a bounded error vector, M is a rank- k matrix and \mathcal{A} is a linear measurement operator that satisfies $2k$ -RIP with constant δ_{2k} (assume $\delta_{2k} < 1/3$). Let X^{t+1} be the $t+1$ -th step iterate of SVP, then the following holds:*

$$\|\mathcal{A}(X^{t+1}) - b\|_2^2 \leq \|\mathcal{A}(M) - b\|_2^2 + 2\delta_{2k}\|\mathcal{A}(X^t) - b\|_2^2.$$

In our analysis, we heavily use the following two results. The first result is the well-known Bernstein's inequality.

Lemma A.2. [Bernstein's inequality] *Let X_1, X_2, \dots, X_n be independent random variables. Also, let $|X_i| \leq L \in \mathbb{R} \forall i$ w.p. 1. Then, we have the following inequality:*

$$\mathbb{P}\left[\left|\sum_{i=1}^n X_i - \sum_{i=1}^n \mathbb{E}[X_i]\right| > t\right] \leq 2 \exp\left(\frac{-t^2/2}{\sum_{i=1}^n \text{Var}(X_i) + Lt/3}\right). \quad (22)$$

The second result is a restatement of Theorem 3.1 from [12].

Theorem A.3. (Restatement of Theorem 3.1 from [12]) *Suppose M is an incoherent rank- k matrix and let p, Ω be as in Theorem 2.5. Further, let M_k be the best rank- k approximation of $\frac{1}{p}P_\Omega(M)$. Then, w.h.p. we have:*

$$\|M - M_k\|_2 \leq C \sqrt{\frac{k}{p\sqrt{mn}}} \|M\|_F. \quad (23)$$

Remark: Note that Theorem 3.1 from [12] holds only for $T_r(P_\Omega(M))$ where $T_r(P_\Omega(M))$ is a trimmed version of $P_\Omega(M)$ obtained by setting all rows and columns of $P_\Omega(M)$ with too many observed entries to zero. However, using standard Chernoff bound we can argue that for our choice of p , none of the rows and columns of $P_\Omega(M)$ have too many observed entries and hence $T_r(P_\Omega(M)) = P_\Omega(M)$, whp.

B Matrix Sensing: Proofs

The following is an alternate characterization of RIP that we use heavily in our proofs. At a conceptual level, it says that if \mathcal{A} satisfies RIP, then it also preserves inner-product between any two rank- k matrices (upto some additive error).

Lemma B.1. *Suppose $\mathcal{A}(\cdot)$ satisfies $2k$ -RIP with constant δ_{2k} . Then, for any $U_1, U_2 \in \mathbb{R}^{m \times k}$ and $V_1, V_2 \in \mathbb{R}^{n \times k}$, we have the following:*

$$\left| \left\langle \mathcal{A}(U_1 V_1^\dagger), \mathcal{A}(U_2 V_2^\dagger) \right\rangle - \text{Tr}(U_2^\dagger U_1 V_1^\dagger V_2) \right| \leq 3\delta_{2k} \|U_1 V_1^\dagger\|_F \|U_2 V_2^\dagger\|_F \quad (24)$$

Proof. Consider the matrices $X_1 \stackrel{\text{def}}{=} U_1 V_1^T$, $X_2 \stackrel{\text{def}}{=} U_2 V_2^T$ and $X = X_1 + X_2$. Since the rank of X is at most $2k$, we obtain the following using the RIP of \mathcal{A} :

$$(1 - \delta) \|U_1 V_1^T + U_2 V_2^T\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1 + \delta) \|U_1 V_1^T + U_2 V_2^T\|_F^2.$$

Concentrating on the second inequality, we obtain

$$\begin{aligned}
& \sum_i (\text{Tr}(A_i U_1 V_1^T) + \text{Tr}(A_i U_2 V_2^T))^2 \leq (1 + \delta) \left(\|U_1 V_1^T\|_F^2 + \|U_2 V_2^T\|_F^2 + \text{Tr}(U_1 V_1^T V_2 U_2^T) \right) \\
& \stackrel{(\zeta_1)}{\Rightarrow} \sum_i \text{Tr}(A_i U_1 V_1^T) \text{Tr}(A_i U_2 V_2^T) - \text{Tr}(U_1 V_1^T V_2 U_2^T) \leq \delta \left(\|U_1 V_1^T\|_F^2 + \|U_2 V_2^T\|_F^2 + \text{Tr}(U_1 V_1^T V_2 U_2^T) \right) \\
& \stackrel{(\zeta_2)}{\Rightarrow} \sum_i \text{Tr}(A_i U_1 V_1^T) \text{Tr}(A_i U_2 V_2^T) - \text{Tr}(U_1 V_1^T V_2 U_2^T) \leq \delta \left(\|U_1 V_1^T\|_F^2 + \|U_2 V_2^T\|_F^2 + \|U_1 V_1^T\|_F \|U_2 V_2^T\|_F \right)
\end{aligned} \tag{25}$$

where (ζ_1) follows from the fact that X_1 and X_2 are rank- k matrices and hence $\mathcal{A}(\cdot)$ satisfies RIP w.r.t. those matrices and (ζ_2) follows from the fact that $\text{Tr}(U_1 V_1^T V_2 U_2^T) \leq \|U_1 V_1^T\|_F \|U_2 V_2^T\|_F$. Note that if we replace $U_1 V_1^T$ by $\lambda U_1 V_1^T$ and $U_2 V_2^T$ by $\frac{1}{\lambda} U_2 V_2^T$ in (25) for some non-zero $\lambda \in \mathbb{R}$, the LHS of (25) does not change where as the RHS of (25) changes. Optimizing the RHS w.r.t. λ , we obtain

$$\sum_i \text{Tr}(A_i U_1 V_1^T) \text{Tr}(A_i U_2 V_2^T) - \text{Tr}(U_1 V_1^T V_2 U_2^T) \leq 3\delta \|U_1 V_1^T\|_F \|U_2 V_2^T\|_F.$$

A similar argument proves the other side of the inequality. This proves the lemma. \square

Proof of Lemma 4.5. We first show that the update (6) reduces to:

$$\sum_{q=1}^k \left(\sum_{i=1}^s A_i u_p^{(t)} u_q^{(t)\dagger} A_i^\dagger \right) \hat{v}_q^{(t+1)} = \sum_{q=1}^k \left(\sum_{i=1}^s A_i u_p^{(t)} u_q^{*\dagger} A_i^\dagger \right) v_q^* \quad \forall p \in [k]. \tag{26}$$

Let $Err(V) \stackrel{\text{def}}{=} \sum_i (\text{Tr}(A_i M) - \text{Tr}(A_i U^{(t)} V^\dagger))^2$. Since $\hat{V}^{(t+1)}$ minimizes $E(V)$, we have $\nabla_V E(\hat{V}^{(t+1)}) = 0$.

$$\begin{aligned}
& \nabla_{v_p} Err(\hat{V}^{(t+1)}) = 0 \\
& \Rightarrow \sum_{i=1}^s \left(\sum_{l=1}^k v_q^{(t)\dagger} A_i u_q^{(t)} - \sum_{l=1}^k \sigma_q^* v_q^{*\dagger} A_i u_q^* \right) A_i u_p = 0 \\
& \Rightarrow \sum_{l=1}^k \sum_{i=1}^s A_i u_p \left(v_q^{(t+1)\dagger} A_i u_q^{(t)} \right) = \sum_{l=1}^k \sum_{i=1}^s A_i u_p \left(\sigma_q^* v_q^{*\dagger} A_i u_q^* \right) \\
& \Rightarrow \sum_{l=1}^k \sum_{i=1}^s A_i u_p \left(u_q^{(t)\dagger} A_i^\dagger v_q^{(t+1)} \right) = \sum_{l=1}^k \sum_{i=1}^s A_i u_p \left(u_q^{*\dagger} A_i^\dagger \sigma_q^* v_q^* \right) \\
& \Rightarrow \sum_{l=1}^k \left(\sum_{i=1}^s A_i u_p u_q^{(t)\dagger} A_i^\dagger \right) v_q^{(t+1)} = \sum_{l=1}^k \left(\sum_{i=1}^s A_i u_p u_q^{*\dagger} A_i^\dagger \right) \sigma_q^* v_q^*
\end{aligned}$$

Define

$$S = \begin{bmatrix} \sigma_1^* I_n & \cdots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \cdots & \sigma_k^* I_n \end{bmatrix}, \quad v^* = \begin{bmatrix} v_1^* \\ \vdots \\ v_k^* \end{bmatrix}, \quad \text{and} \quad \hat{v}_1^{(t+1)} = \begin{bmatrix} \hat{v}_1^{(t+1)} \\ \vdots \\ \hat{v}_k^{(t+1)} \end{bmatrix}.$$

Then,

$$\begin{aligned}\hat{v}_1^{(t+1)} &= B^{-1}CSv^* \\ &= DSv^* - B^{-1}(BD - C)Sv^*\end{aligned}$$

where inverting B is valid since the minimum singular value of B is strictly positive (please refer Lemma B.2). Considering the p^{th} block of $\hat{v}^{(t)}$, we obtain

$$\begin{aligned}\hat{v}_p^{(t+1)} &= \left(\sum_q \langle u_p^{(t)}, u_q^* \rangle \sigma_q^* v_q^* \right) - (B^{-1}(BD - C)Sv^*)_p \\ &= \left(\sum_q \sigma_q^* v_q^* u_q^{*\dagger} \right) u_p^{(t)} - (B^{-1}(BD - C)Sv^*)_p.\end{aligned}$$

This gives us the following equation for $\hat{V}^{(t)}$:

$$\hat{V}^{(t+1)} = V^* \Sigma^* U^{*\dagger} U^{(t)} - F$$

where $F = \begin{bmatrix} (B^{-1}(BD - C)Sv^*)_1 & (B^{-1}(BD - C)Sv^*)_2 & \dots & (B^{-1}(BD - C)Sv^*)_k \end{bmatrix}$.

Hence Proved. \square

B.1 Rank-1 Matrix Sensing: Proofs

Proof of Lemma 4.3. Using definition of the spectral norm:

$$\|B^{-1}(\langle u^*, u^t \rangle B - C)v^*\| \leq \|B^{-1}\|_2 \cdot \|\langle u^*, u^t \rangle B - C\|_2 \cdot \|v^*\|_2. \quad (27)$$

Consider $B = \sum_i A_i u^t (u^t)^\dagger A_i^\dagger$. Now, smallest eigenvalue of B , i.e., $\lambda_{\min}(B)$ is given by:

$$\begin{aligned}\lambda_{\min}(B) &= \min_{\|z\|=1} z^\dagger B z = \min_{\|z\|=1} \sum_i z^\dagger A_i u^t (u^t)^\dagger A_i^\dagger z = \min_{\|z\|=1} \sum_i \text{Tr}(A_i u^t z^\dagger) \text{Tr}(A_i u^t z^\dagger), \\ &= \min_{\|z\|=1} \langle \mathcal{A}(u^t z^\dagger), \mathcal{A}(u^t z^\dagger) \rangle \geq 1 - 3\delta_2,\end{aligned} \quad (28)$$

where the last inequality follows using Lemma B.1. Using (28),

$$\|B^{-1}\|_2 \leq \frac{1}{1 - 3\delta_2}. \quad (29)$$

Now, consider $G = \langle u^*, u^t \rangle B - C = \sum_i A_i (\langle u^*, u^t \rangle u^t (u^t)^\dagger - u^t (u^*)^\dagger) A_i^\dagger = \sum_i A_i u^t (\langle u^*, u^t \rangle u^t - u^*)^\dagger A_i^\dagger$. Using definition of the spectral norm:

$$\begin{aligned}\|G\|_2 &= \max_{\|z\|=1, \|y\|=1} z^\dagger G y, \\ &= \max_{\|z\|=1, \|y\|=1} \sum_i z^\dagger A_i u^t (\langle u^*, u^t \rangle u^t - u^*)^\dagger A_i^\dagger y, \\ &= \max_{\|z\|=1, \|y\|=1} \langle \mathcal{A}(u^t z^\dagger), \mathcal{A}((\langle u^*, u^t \rangle u^t - u^*) y^\dagger) \rangle, \\ &\leq 3\delta_2 \sqrt{1 - \langle u^t, u^* \rangle^2},\end{aligned} \quad (30)$$

where the last inequality follows by using Lemma B.1 and the fact that $\langle u^t, (\langle u^*, u^t \rangle u^t - u^*) \rangle = 0$.

Lemma now follows using (27), (29), (30). \square

B.2 Rank- k Matrix Sensing

Proof of Lemma 4.4. Since \widehat{U}^t and \widetilde{U}^t have full rank and span the same subspace, there exists a $k \times k$, full rank matrix R such that $\widehat{U}^t = \widetilde{U}^t R = U^t R_U^t R$. We have:

$$\left\| \mathcal{A} \left(\widehat{U}^t V^\dagger \right) - b \right\|_2 = \left\| \mathcal{A} \left(U^t \left(V \left(R_U^t R \right)^\dagger \right)^\dagger \right) - b \right\|_2 \geq \left\| \mathcal{A} \left(U^t \left(\widetilde{V}^{t+1} \right)^\dagger \right) - b \right\|_2$$

with equality holding in the last step for $V = \widetilde{V}^{t+1} \left(\left(R_U^t R \right)^\dagger \right)^{-1}$. The proof of Theorem 4.2 shows that \widetilde{V}^{t+1} is unique and has full rank (since $\text{dist} \left(\widetilde{V}^{t+1}, V^* \right) < 1$). This means that \widehat{V}^{t+1} is also unique and is equal to $\widetilde{V}^{t+1} \left(\left(R_U^t R \right)^\dagger \right)^{-1}$. This shows that $\text{Span} \left(\widehat{V}^{t+1} \right) = \text{Span} \left(\widetilde{V}^{t+1} \right)$ and that both \widehat{V}^{t+1} and \widetilde{V}^{t+1} have full rank. \square

Lemma B.2. *Let linear measurement \mathcal{A} satisfy RIP for all $2k$ -rank matrices and let $b = \mathcal{A}(M)$ with $M \in \mathbb{R}^{m \times n}$ being a rank- k matrix. Let δ_{2k} be the RIP constant for rank $2k$ -matrices. Then, we have the following bound on the minimum singular value of B :*

$$\sigma_{\min}(B) \geq 1 - \delta_{2k}. \quad (31)$$

Proof. Select any $w \in \mathbb{R}^{nk}$ such that $\|w\|_2 = 1$. Let

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}$$

where each $w_p \in \mathbb{R}^n$. Also denote $W \stackrel{\text{def}}{=} [w_1 w_2 \cdots w_k] \in \mathbb{R}^{n \times k}$, i.e., $w = \text{vec}(W)$.

We have,

$$\begin{aligned} w^\dagger B w &= \sum_{p,q=1}^k w_p^\dagger B_{pq} w_q = \sum_{p,q=1}^k w_p^\dagger \left(\sum_{i=1}^d A_i u_p^t (u_q^t)^\dagger A_i^\dagger \right) w_q = \sum_{i=1}^d \sum_{p,q=1}^k w_p^\dagger A_i u_p^t (u_q^t)^\dagger A_i^\dagger w_q \\ &= \sum_{i=1}^d \left(\sum_{p=1}^k w_p^\dagger A_i u_p^t \right) \left(\sum_{q=1}^k w_q^\dagger A_i u_q^t \right) = \sum_{i=1}^d \text{Tr} \left(A_i U^t W^\dagger \right)^2. \end{aligned}$$

Now, using RIP (see Definition 2.1) along with the above equation, we get:

$$w^\dagger B w = \sum_{i=1}^d \text{Tr} \left(A_i U^t W^\dagger \right)^2 \geq (1 - \delta) \left\| U^t W^\dagger \right\|_F^2 = (1 - \delta_{2k}) \|W\|_F^2 = (1 - \delta_{2k}) \|w\|^2 = (1 - \delta_{2k}).$$

Since w was arbitrary, this proves the lemma. \square

Proof of Lemma 4.6. Note that,

$$\begin{aligned} \|F(\Sigma^*)^{-1}\|_2 &\leq \|F(\Sigma^*)^{-1}\|_F = \|B^{-1} (BD - C) v^*\|_2 \\ &\leq \|B^{-1}\|_2 \|(BD - C)\|_2 \|v^*\|_2 \\ &\leq \frac{\sqrt{k}}{1 - \delta_{2k}} \|(BD - C)\|_2 \end{aligned} \quad (32)$$

where the last step follows from Lemma B.2. Now we need to bound $\|(BD - C)\|_2$. Choose any $w, z \in \mathbb{R}^{nk}$ such that $\|w\|_2 = \|z\|_2 = 1$. As in Lemma B.2, define the following components of w and z :

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix}$$

where each $w_p, z_p \in \mathbb{R}^n$ and $W \stackrel{\text{def}}{=} [w_1 w_2 \cdots w_k]$ and $Z \stackrel{\text{def}}{=} [z_1 z_2 \cdots z_k] \in \mathbb{R}^{n \times k}$. We have,

$$w^\dagger (BD - C) z = \sum_{p,q=1}^k w_p^\dagger (BD - C)_{pq} z_q$$

We calculate $(BD - C)_{pq}$ as follows:

$$\begin{aligned} (BD - C)_{pq} &= \sum_{l=1}^k B_{pl} D_{lq} - C_{pq} = \left(\sum_{l=1}^k B_{pl} \langle u_l^t, u_q^* \rangle \mathbb{I}_{n \times n} \right) - C_{pq} = \left(\sum_{l=1}^k u_q^{*\dagger} u_l^t \sum_{i=1}^d A_i u_p^t (u_l^t)^\dagger A_i^\dagger \right) - C_{pq} \\ &= \sum_{i=1}^d A_i u_p^t u_q^{*\dagger} \sum_{l=1}^k u_l^t (u_l^t)^\dagger A_i^\dagger - \sum_{i=1}^d A_i u_p^t (u_q^*)^\dagger A_i^\dagger = \sum_{i=1}^d A_i u_p^t u_q^{*\dagger} \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) A_i^\dagger. \end{aligned}$$

So we have,

$$\begin{aligned} w^\dagger (BD - C) z &= \sum_{p,q=1}^k w_p^\dagger \sum_{i=1}^d A_i u_p^t u_q^{*\dagger} \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) A_i^\dagger z_q = \sum_{i=1}^d \sum_{p,q=1}^k w_p^\dagger A_i u_p^t u_q^{*\dagger} \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) A_i^\dagger z_q \\ &= \sum_{i=1}^d \text{Tr} \left(A_i U^t W^\dagger \right) \text{Tr} \left(A_i \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) U^* Z^\dagger \right) \\ &\stackrel{(\zeta_1)}{\leq} \text{Tr} \left(U^{*\dagger} \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) U^t W^\dagger Z \right) + \delta_{2k} \left\| U^t W^\dagger \right\|_F \left\| \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) U^* Z^\dagger \right\|_F \\ &\stackrel{(\zeta_2)}{\leq} \delta_{2k} \|W\|_F \sqrt{\left\| (U^*)^\dagger (U^t (U^t)^\dagger - \mathbb{I}_{n \times n})^2 U^* \right\|_F} \|Z^\dagger Z\|_F \\ &\stackrel{(\zeta_3)}{\leq} \delta_{2k} \sqrt{k} \cdot \text{dist}(U^t, U^*), \end{aligned}$$

where (ζ_1) follows from the fact that \mathcal{A} satisfies $2k$ -RIP and Lemma B.1, (ζ_2) follows from the fact that $(U^t (U^t)^\dagger - \mathbb{I}_{n \times n}) U^t = 0$, (ζ_3) follows from the following: $\|W\|_F = \|w\|_2 = 1$, $\|Z^\dagger Z\|_F \leq \|Z\|_F^2 = 1$ and finally: $\left\| (U^t (U^t)^\dagger - \mathbb{I}_{n \times n}) U^* \right\|_F \leq \sqrt{k} \left\| (U^t (U^t)^\dagger - \mathbb{I}_{n \times n}) U^* \right\|_2$.

Since w and z were arbitrary unit vectors, we can conclude that $\|BD - C\|_2 \leq \delta_{2k} \sqrt{k} \cdot \text{dist}(U^t, U^*)$. Plugging this bound in (32) proves the lemma. \square

Proof of Lemma 4.7. Note that $\|\Sigma^*(R^{(t+1)})^{-1}\|_2 \leq \frac{\sigma_1^*}{\sigma_{\min}(R^{(t+1)})}$. Now,

$$\begin{aligned}
\sigma_{\min}(R^{(t+1)}) &= \min_{z, \|z\|_2=1} \|R^{(t+1)}z\|_2 = \min_{z, \|z\|_2=1} \|V^{(t+1)}R^{(t+1)}z\|_2, \\
&= \min_{z, \|z\|_2=1} \|V^*\Sigma^*U^{*\dagger}U^{(t)}z - Fz\|_2, \\
&\geq \min_{z, \|z\|_2=1} \|V^*\Sigma^*U^{*\dagger}U^{(t)}z\|_2 - \|Fz\|_2, \\
&\geq \min_{z, \|z\|_2=1} \|V^*\Sigma^*U^{*\dagger}U^{(t)}z\|_2 - \|F\|_2, \\
&\geq \sigma_k^* \sigma_{\min}(U^{*\dagger}U^{(t)}) - \|F\|_2, \\
&\geq \sigma_k^* \sqrt{1 - \|U_{\perp}^{*\dagger}U^{(t)}\|_2^2} - \sigma_1^* \|F(\Sigma^*)^{-1}\|_2, \\
&= \sigma_k^* \sqrt{1 - \text{dist}(U^*, U^{(t)})^2} - \sigma_1^* \|F(\Sigma^*)^{-1}\|_2.
\end{aligned} \tag{33}$$

Lemma now follows using above inequality with Lemma 4.6. \square

B.3 Noisy Matrix Sensing: Proofs

We now consider an extension of the matrix sensing problem where measurements can be corrupted arbitrarily using a bounded noise. That is, we observe $b = \mathcal{A}(M + N)$, where N is the noise matrix. For this noisy case as well, we show that AltMinSense recovers M upto an additive approximation depending on the Frobenius norm of N .

Theorem B.3. *Let M and $\mathcal{A}(\cdot)$ be as defined in Theorem 2.2. Suppose, AltMinSense algorithm (Algorithm 1) is supplied inputs \mathcal{A} , $b = \mathcal{A}(M + N)$, where N is the noise matrix s.t. $\|N\|_F < \frac{1}{100}\sigma_k^*$. Then, after $T = 4\log(2/\epsilon)$ steps, iterates \hat{U}^T , \hat{V}^T of AltMinSense satisfy:*

$$\text{dist}(\hat{V}^T, V^*) \leq \frac{10\|N\|_F}{\sigma_k^*} + \epsilon, \quad \text{dist}(\hat{U}^T, U^*) \leq \frac{10\|N\|_F}{\sigma_k^*} + \epsilon.$$

See Definition 4.1 for definition of $\text{dist}(U, W)$.

Proof. At a high level, our proof for noisy case follows closely, the exact case proof given in Section 4. That is, we show that the update of AltMinSense algorithm is similar to power-method type update but with two errors terms: one due to incomplete measurements and another due to the noise matrix.

Similar to our proof for sensing problem (Section 4), we analyze QR-decomposition based updates. That is,

$$\begin{aligned}
\hat{U}^t &= U^t R_U^t \quad (\text{QR decomposition}), \\
\hat{V}^{t+1} &= \underset{V}{\text{argmin}} \|\mathcal{A}(U^t V^\dagger) - b\|_2^2, \\
\hat{V}^{t+1} &= V^{t+1} R_V^{t+1}. \quad (\text{QR decomposition})
\end{aligned}$$

Similar to Lemma 4.5, we can re-write the above given update equation as:

$$\begin{aligned}\widehat{V}^{t+1} &= V^* \Sigma^* (U^*)^\dagger U^t - F + V^N \Sigma^N (U^N)^\dagger U^{(t)} - G, \\ V^{t+1} &= \widehat{V}^{t+1} (R^{(t+1)})^{-1},\end{aligned}\tag{34}$$

where, F is the error matrix and is as defined in (8) and G is the error matrix due to noise and is given by:

$$G \stackrel{\text{def}}{=} [(B^{-1} (BD^N - C^N) S^N v^N)_1 \quad \dots \quad (B^{-1} (BD^N - C^N) S^N v^N)_k], \tag{35}$$

where B , C and D defined in the previous section (See (7)) and C^N and D^N are defined below:

$$C^N \stackrel{\text{def}}{=} \begin{bmatrix} C_{11}^N & \dots & C_{1m}^N \\ \vdots & \ddots & \vdots \\ C_{k1}^N & \dots & C_{km}^N \end{bmatrix}, \quad D^N \stackrel{\text{def}}{=} \begin{bmatrix} D_{11}^N & \dots & D_{1m}^N \\ \vdots & \ddots & \vdots \\ D_{k1}^N & \dots & D_{km}^N \end{bmatrix},$$

with $C_{pq}^N \stackrel{\text{def}}{=} \sum_{i=1}^d A_i u_p^{(t)} (u_q^N)^\dagger A_i^\dagger$ and $D_{pq}^N \stackrel{\text{def}}{=} \langle u_p^{(t)}, u_q^N \rangle \mathbb{I}_{n \times n}$. Also,

$$S^N = \begin{bmatrix} \sigma_1^N I_n & \dots & 0_n \\ \vdots & \ddots & \vdots \\ 0_n & \dots & \sigma_N^N I_n \end{bmatrix}, \quad v^N = \begin{bmatrix} v_1^N \\ \vdots \\ v_k^N \end{bmatrix}.$$

Now, multiplying (34) with V_\perp^* , we get:

$$V_\perp^{*\dagger} V^{t+1} = (V_\perp^{*\dagger} V^N \Sigma^N U^N \dagger U^{(t)} - V_\perp^{*\dagger} F - V_\perp^{*\dagger} G) R^{(t+1)-1}.$$

That is,

$$\begin{aligned}\text{dist}(V^*, V^{t+1}) &= \|V_\perp^{*\dagger} V^{t+1}\|_2 \\ &\leq (\|V^N \Sigma^N (U^N)^\dagger U^{(t)}\|_2 + \|F\|_2 + \|G\|_2) \|(R^{(t+1)})^{-1}\|_2, \\ &\leq (\sigma_1^N + \|F(\Sigma^*)^{-1}\|_2 \|\Sigma^*\|_2 + \|G\|_2) \|(R^{(t+1)})^{-1}\|_2, \\ &\leq \left(\sigma_1^N + \frac{\sigma_1^* \delta_{2k} k}{1 - \delta_{2k}} \text{dist}(U^t, U^*) + \|G\|_2 \right) \|(R^{(t+1)})^{-1}\|_2,\end{aligned}\tag{36}$$

where the last inequality follows using Lemma 4.6.

Now, we break down the proof in the following two steps:

- Bound $\|G\|_2$ (Lemma B.4, analogous to Lemma 4.6)
- Bound $\|(R^{(t+1)})^{-1}\|_2$ (Lemma B.5, similar to Lemma 4.7)

Later in this section, we provide the above mentioned lemmas and their detailed proof.

Now, by assumption, $\sigma_1^N \leq \|N\|_F \leq \sigma_k^*$. Also, as $\delta_{2k} \leq 1/2$, $\frac{1}{1-\delta_{2k}} \leq 2$. Finally, assume $\text{dist}(V^*, V^{t+1}) \geq \max(10 \cdot \frac{\sigma_1^N}{\sigma_k^*}, 10 \frac{\|N\|_F}{\sigma_1^*})$. Using these observations and lemmas B.4, B.5 along with (36), we get:

$$\text{dist}(V^*, V^{t+1}) \leq \frac{0.5 \text{dist}(U^*, U^t)}{\sqrt{1 - \text{dist}(U^t, U^*)^2} - 0.5 \text{dist}(U^*, U^t)}.\tag{37}$$

As, U^0 is obtained using SVD of $\sum_i A_i b_i$. Hence, using Lemma A.1, we have:

$$\begin{aligned}
& \|\mathcal{A}(U^0 \Sigma^0 V^0 - U^* \Sigma^* (V^*)^\dagger)\|_2^2 \leq 0.5 \|\mathcal{A}(N)\|_2^2 + 4\delta_{2k} \|\mathcal{A}(U^* \Sigma^* (V^*)^\dagger)\|_2^2, \\
& \Rightarrow \|U^0 \Sigma^0 V^0 - U^* \Sigma^* (V^*)^\dagger\|_F^2 \leq \|N\|_F^2 + 4\delta_{2k}(1 + \delta_{2k}) \|\Sigma^*\|_F^2, \\
& \Rightarrow (\sigma_k^*)^2 \|(U^0 (U^0)^\dagger - I)U^*\|_F^2 \leq \|N\|_F^2 + 4\delta_{2k}(1 + \delta_{2k})k(\sigma_1^*)^2, \\
& \Rightarrow \text{dist}(U^0, U^*) \leq \|(U^0 (U^0)^\dagger - I)U^*\|_F^2 \leq \frac{\|N\|_F^2}{(\sigma_k^*)^2} + 6\delta_{2k}k \left(\frac{\sigma_1^*}{\sigma_k^*}\right)^2 < \frac{1}{2},
\end{aligned}$$

where last inequality follows using $\frac{\|N\|_F}{\sigma_k^*} < 1/100$.

Theorem now follows using above equation with (37). \square

Lemma B.4. *Let linear measurement \mathcal{A} satisfy RIP for all $2k$ -rank matrices and let $b = \mathcal{A}(M + N)$ with $M \in \mathbb{R}^{m \times n}$ being a rank- k matrix and let $N = U^N \Sigma^N (V^N)^\dagger$. Let δ_{2k} be the RIP constant for rank $2k$ -matrices. Then, we have the following bound on $\|G\|_2$:*

$$\|G\|_2 \leq \frac{\delta_{2k} \|N\|_F}{1 - \delta_{2k}}. \quad (38)$$

Proof. Note that,

$$\begin{aligned}
\|G\|_2 & \leq \|G\|_F = \|B^{-1}(BD^N - C^N)S^N v^N\|_2 \\
& \leq \|B^{-1}\|_2 \|(BD^N - C^N)S^N\|_2 \|S^N v^N\|_2 \leq \frac{\sqrt{k}}{1 - \delta_{2k}} \|(BD^N - C^N)S^N\|_2, \quad (39)
\end{aligned}$$

where the last inequality follows using Lemma B.2 and the fact that $\|V^N\|_F = \sqrt{k}$. Now let $w = [w_1^\dagger \ w_2^\dagger \ \dots \ w_k^\dagger]^\dagger \in \mathbb{R}^{nk}$ and $z = [z_1^\dagger \ z_2^\dagger \ \dots \ z_n^\dagger]^\dagger \in \mathbb{R}^{n^2}$ be any two arbitrary vectors such that $\|w\|_2 = \|z\|_2 = 1$. Then,

$$\begin{aligned}
w^\dagger (BD^N - C^N) S^N z & = \sum_{p=1}^k \sum_{q=1}^n w_p^\dagger \sum_{i=1}^d A_i u_p^t u_q^{N\dagger} \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) A_i^\dagger \sigma_q^N z_q \\
& = \sum_{i=1}^d \sum_{p=1}^k \sum_{q=1}^n w_p^\dagger A_i u_p^t u_q^{N\dagger} \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) A_i^\dagger \sigma_q^N z_q \\
& = \sum_{i=1}^d \left(\sum_{p=1}^k w_p^\dagger A_i u_p^t \right) \left(\sum_{q=1}^n \sigma_q^N z_q^\dagger A_i \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) u_q^N \right) \\
& = \sum_{q=1}^n \left(\sum_{i=1}^d \text{Tr} \left(A_i U W^\dagger \right) \text{Tr} \left(A_i \left(U^t (U^t)^\dagger - \mathbb{I}_{n \times n} \right) u_q^N \sigma_q^N z_q^\dagger \right) \right).
\end{aligned}$$

Now, using RIP, we get:

$$\begin{aligned}
w^\dagger (BD^N - C^N) z &\leq \sum_{q=1}^n u_q^{N\dagger} \left(U^t U^{t\dagger} - \mathbb{I}_{n \times n} \right) U^t W^\dagger \sigma_q^N z_q + \delta_{2k} \left\| U^t W^\dagger \right\|_F \left\| \left(U^t U^{t\dagger} - \mathbb{I}_{n \times n} \right) u_q^N \sigma_q^N z_q^\dagger \right\|_F \\
&\leq \sum_{q=1}^n \delta_{2k} \|W^\dagger\|_F \|(U^t(U^t)^\dagger - \mathbb{I}_{n \times n})u_q^N\|_2 \|\sigma_q^N z_q\|_2, \\
&\leq \delta_{2k} \sum_{q=1}^n \|u_q^N\|_2 \|\sigma_q^N z_q\|_2 = \delta_{2k} \sum_{q=1}^n \sigma_q^N \|z_q\|_2, \\
&\leq \delta_{2k} \sqrt{\sum_{q=1}^n \|z_q\|_2^2} \sqrt{\sum_{q=1}^n (\sigma_q^N)^2} \leq \delta_{2k} \|N\|_F.
\end{aligned}$$

This finishes the proof. \square

Lemma B.5. *Assuming conditions of Lemma B.4, we have the following bound on the minimum singular value of $R^{(t)}$:*

$$\sigma_{\min} \left(R^{(t+1)} \right) \geq \sigma_k^* \sqrt{1 - \text{dist}(U^t, U^*)^2} - \sigma_1^N - \|F\|_2 - \|G\|_2.$$

Proof. Similar to the proof of Lemma 4.7, we have the following set of inequalities:

$$\begin{aligned}
\sigma_{\min} \left(R^{(t+1)} \right) &= \min_{\|z\|_2=1} \left\| R^{(t+1)} z \right\|_2 = \min_{\|z\|_2=1} \left\| V^{(t)} R^{(t+1)} z \right\|_2 \\
&= \min_{\|z\|_2=1} \left\| V^* \Sigma^* U^{*\dagger} U^t z + V^N \Sigma^N (U^N)^\dagger U^{(t)} z - Fz - Gz \right\|_2 \\
&\geq \min_{\|z\|_2=1} \left\| V^* \Sigma^* U^{*\dagger} U^t z \right\|_2 - \left\| V^N \Sigma^N (U^N)^\dagger \right\|_2 - \|F\|_2 - \|G\|_2 \\
&\geq \sigma_k^* \min_{\|z\|_2=1} \left\| U^{*\dagger} U^t z \right\|_2 - \sigma_1^N - \|F\|_2 - \|G\|_2 \\
&\geq \sigma_k^* \sqrt{1 - \|U_\perp^{*\dagger} U\|_2^2} - \sigma_1^N - \|F\|_2 - \|G\|_2 \\
&= \sigma_k^* \sqrt{1 - \text{dist}(U^t, U^*)^2} - \sigma_1^N - \|F\|_2 - \|G\|_2.
\end{aligned}$$

This proves the lemma. \square

B.4 Stagewise Alternating Minimization for Matrix Sensing: Proofs

Proof of Lemma 6.1. As the initial point of the i -th stage is obtained by one step of SVP [10], using Lemma A.1, we obtain:

$$\left\| M - \widehat{U}_{1:i}^0 (\widehat{V}_{1:i}^0)^\dagger \right\|_F^2 \leq \sum_{j=i+1}^k (\sigma_j^*)^2 + 2\delta_{2k} \|M - \widehat{U}_{1:i-1}^T V_{1:i-1}^T\|_F^2.$$

Now, by assumption over the $(i-1)$ -th stage error (this assumption follows from the inductive hypothesis in proof of Theorem 2.3),

$$\left\| M - \widehat{U}_{1:i}^0 (\widehat{V}_{1:i}^0)^\dagger \right\|_F^2 \leq \sum_{j=i+1}^k (\sigma_j^*)^2 + 2\delta_{2k} 16k (\sigma_i^*)^2.$$

Lemma now follows by setting $\delta_{2k} \leq \frac{1}{3200k}$. \square

Proof of Lemma 6.2. For our proof, we consider two cases: a) $\frac{\sigma_i^*}{\sigma_{i+1}^*} < 5\sqrt{k}$, b) $\frac{\sigma_i^*}{\sigma_{i+1}^*} \geq 5\sqrt{k}$.

Case (a): In this case, using monotonicity of the AltMin algorithm directly gives error bound. That is, $\|M - \widehat{U}_{1:i}^T(\widehat{V}_{1:i}^T)^\dagger\|_F^2 \leq \|M - \widehat{U}_{1:i}^0 V_{1:i}^0\|_F^2 \leq k(\sigma_{i+1}^*)^2 + \frac{25k}{100}(\sigma_{i+1}^*)^2$.

Case (b): At a high level, if $\frac{\sigma_i^*}{\sigma_{i+1}^*} \geq 5\sqrt{k}$ then $U_{1:i}^0$ is “close” to $U_{1:i}^*$ and hence the error bound follows by using an analysis similar to the noisy case. Note that σ_{i+1}^* being small implies that the “noise” is small. See Lemma B.6 for a formal proof of this case. \square

Lemma B.6. Assume conditions given in Theorem 2.3 are satisfied and let $\frac{\sigma_i^*}{\sigma_{i+1}^*} \geq 5\sqrt{k}$. Also, let

$$\left\|M - \widehat{U}_{1:i}^0(\widehat{V}_{1:i}^0)^\dagger\right\|_F^2 \leq \sum_{j=i+1}^k (\sigma_j^*)^2 + \frac{1}{100}(\sigma_i^*)^2.$$

Then, $U_{1:i}^T, V_{1:i}^T$ satisfy:

$$\|M - \widehat{U}_{1:i}^T V_{1:i}^T\|_F^2 \leq \max(\epsilon, 16k(\sigma_{i+1}^*)^2),$$

Proof. We first show that if σ_i and σ_{i+1} have large gap then $\forall t$, the t^{th} iterate of the i -th stage, $\widehat{U}_{1:i}^t$ is close to $U_{1:i}^*$. Let U_\perp^t be a basis of the subspace orthogonal to $\widehat{U}_{1:i}^t$.

$$\begin{aligned} \|(U_\perp^t)^\dagger (M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger)\|_2 &= \|(U_\perp^t)^\dagger M\|_2 \geq \|(U_\perp^t)^\dagger U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger\|_2 - \|(U_\perp^t)^\dagger U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger\|_2, \\ &\geq \sigma_i^* \|(U_\perp^t)^\dagger U_{1:i}^*\|_2 - \sigma_{i+1}^* \geq \sigma_i^* (\|(U_\perp^t)^\dagger U_{1:i}^*\|_2 - \frac{1}{5\sqrt{k}}). \end{aligned} \quad (40)$$

We also have:

$$\begin{aligned} \|(U_\perp^t)^\dagger (M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger)\|_2^2 &\leq \|M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger\|_F^2 \leq \frac{1}{1 - \delta_{2k}} \left\| \mathcal{A} \left(M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger \right) \right\|_2^2 \\ &\stackrel{(\zeta_1)}{\leq} \frac{1}{1 - \delta_{2k}} \left\| \mathcal{A} \left(M - \widehat{U}_{1:i}^0(\widehat{V}_{1:i}^0)^\dagger \right) \right\|_2^2 \leq \frac{1 + \delta_{2k}}{1 - \delta_{2k}} \|M - \widehat{U}_{1:i}^0(\widehat{V}_{1:i}^0)^\dagger\|_F^2 \\ &\leq \frac{1 + \delta_{2k}}{1 - \delta_{2k}} \left(\sum_{j=i+1}^k (\sigma_j^*)^2 + \frac{1}{100}(\sigma_i^*)^2 \right) \\ &\leq \frac{1 + \delta_{2k}}{1 - \delta_{2k}} \left(k(\sigma_{i+1}^*)^2 + \frac{1}{100}(\sigma_i^*)^2 \right), \end{aligned} \quad (41)$$

where (ζ_1) follows from the fact that lines 5–8 of Algorithm 3 never increases $\left\| \mathcal{A} \left(M - \widehat{U}_{1:i}^t(\widehat{V}_{1:i}^t)^\dagger \right) \right\|_2$. Using (40), (41), and $\frac{\sigma_i^*}{\sigma_{i+1}^*} \geq 5\sqrt{k}$, we obtain the following bound:

$$\|(U_\perp^t)^\dagger U_{1:i}^*\|_2 \leq \frac{1}{2} \quad \forall t. \quad (42)$$

Now, we consider the update equation for \widehat{V}^{t+1} :

$$\widehat{V}^{t+1} = \underset{\widehat{V}}{\operatorname{argmin}} \| \mathcal{A}(\widehat{U}_{1:i}^t \widehat{V} - U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger - U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger) \|_2^2.$$

Note that, the update is same as noisy case with noise matrix $N = U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger$ from (34):

$$\widehat{V}^{t+1} = V_{1:i}^* \Sigma_{1:i}^* (U_{1:i}^*)^\dagger U_{1:i}^t - F + V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t - G, \quad (43)$$

where F and G are given by (8), (35). Multiplying (43) from the left by $V_\perp^\dagger = I - V^{t+1}(V^{t+1})^\dagger$, we obtain:

$$\begin{aligned} 0 &= V_\perp^\dagger \widehat{V}^{t+1} = V_\perp^\dagger \left(V_{1:i}^* \Sigma_{1:i}^* (U_{1:i}^*)^\dagger U_{1:i}^t - F + V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t - G \right) \\ &\Rightarrow V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* (U_{1:i}^*)^\dagger U_{1:i}^t = V_\perp^\dagger \left(F - V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t + G \right) \\ &\Rightarrow \left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* (U_{1:i}^*)^\dagger U_{1:i}^t \right\|_F \leq \|F\|_F + \left\| V_\perp^\dagger V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t \right\|_F + \|G\|_F \\ &\Rightarrow \left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* \right\|_F \leq \frac{1}{\sigma_{\min}((U_{1:i}^*)^\dagger U_{1:i}^t)} \left(\|F\|_F + \left\| V_\perp^\dagger V_{i+1:k}^* \Sigma_{i+1:k}^* (U_{i+1:k}^*)^\dagger U_{1:i}^t \right\|_F + \|G\|_F \right), \quad (44) \end{aligned}$$

where the last inequality follows using the fact that $\sigma_{\min}(A)\|B\|_F \leq \|AB\|_F$. Using Lemma B.4, and a modification of Lemma 4.6, we get:

$$\|F\|_F \leq \delta_{2k} \left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F, \quad \|G\|_F \leq \delta_{2k} \left\| U_\perp^\dagger U_{i+1:k}^* \Sigma_{i+1:k}^* \right\|_F \leq \delta_{2k} \sqrt{k} \sigma_{i+1}. \quad (45)$$

Using (44), (45), and the fact that $\sigma_{\min}(U_\perp^\dagger U_{1:i}^*) = \sqrt{1 - \|U_\perp^\dagger U_{1:i}^*\|_2^2}$,

$$\left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* \right\|_F \leq \frac{2}{\sqrt{3}} \left(\delta_{2k} \left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F + \sqrt{\sum_{j=i+1}^k (\sigma_j^*)^2} + \delta_{2k} \sqrt{k} \sigma_{i+1} \right).$$

Assuming $\left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F > 2\sqrt{\sum_{j=i+1}^k \sigma_j^2}$, we obtain:

$$\left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* \right\|_F \leq \frac{2}{3} \left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F. \quad (46)$$

Using similar analysis, we can show that,

$$\left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F \leq \frac{2}{3} \left\| V_\perp^\dagger V_{1:i}^* \Sigma_{1:i}^* \right\|_F.$$

So after $T \geq 8 \log(k\sigma_i^*)$ iterations, we have:

$$\left\| U_\perp^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F^2 \leq 4 \sum_{j=i+1}^k (\sigma_j^*)^2.$$

Using the above inequality, we now bound the error after $T \geq 8 \log(k\sigma_i^*)$ iterations of the i -th stage:

$$\left\| M - \widehat{U}_{1:i}^T (\widehat{V}_{1:i}^T)^\dagger \right\|_F \leq \left\| U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger - \widehat{U}_{1:i}^T (\widehat{V}_{1:i}^T)^\dagger \right\|_F + \left\| U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger \right\|_F. \quad (47)$$

For the first term, we have:

$$\begin{aligned}
& \left\| U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger - \widehat{U}_{1:i}^T (\widehat{V}_{1:i}^T)^\dagger \right\|_F^2 \\
&= \left\| U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger - U_{1:i}^T (U_{1:i}^T)^\dagger U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger + U_{1:i}^T (U_{1:i}^T)^\dagger U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger - \widehat{U}_{1:i}^T (\widehat{V}_{1:i}^T)^\dagger \right\|_F^2 \\
&= \left\| \left(I - U_{1:i}^T (U_{1:i}^T)^\dagger \right) U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger \right\|_F^2 + \left\| U_{1:i}^T (U_{1:i}^T)^\dagger U_{1:i}^* \Sigma_{1:i}^* (V_{1:i}^*)^\dagger - \widehat{U}_{1:i}^T (\widehat{V}_{1:i}^T)^\dagger \right\|_F^2 \\
&\stackrel{(\zeta_1)}{\leq} \left\| U_{\perp}^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F^2 + \left\| U_{1:i}^T \left(F + G - (U_{1:i}^t)^\dagger U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger \right) \right\|_F^2 \\
&= \left\| U_{\perp}^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F^2 + \left\| F + G - (U_{1:i}^t)^\dagger U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger \right\|_F^2 \\
&= \left\| U_{\perp}^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F^2 + 3 \|F\|_F^2 + 3 \|G\|_F^2 + 3 \left\| U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger \right\|_F^2 \\
&\stackrel{(\zeta_2)}{\leq} (1 + 3\delta_{2k}^2) \left\| U_{\perp}^\dagger U_{1:i}^* \Sigma_{1:i}^* \right\|_F^2 + 3(1 + \delta_{2k}^2) \left\| U_{i+1:k}^* \Sigma_{i+1:k}^* (V_{i+1:k}^*)^\dagger \right\|_F^2 \\
&\leq 8k(\sigma_{i+1}^*)^2,
\end{aligned} \tag{48}$$

where (ζ_1) follows from (43) and (ζ_2) follows from (45). Using (47) and (48), we obtain the following bound:

$$\left\| M - \widehat{U}_{1:i}^T (\widehat{V}_{1:i}^T)^\dagger \right\|_F \leq 4\sqrt{k}\sigma_{i+1}^*. \tag{49}$$

Hence Proved. \square

C Matrix Completion: Proofs

Proof Of Theorem 2.5. Using Theorem 5.1, after $O(\log(1/\epsilon))$ iterations, we get:

$$\text{dist}(U^t, U^*) \leq \epsilon, \quad \text{dist}(V^{t+1}, V^*) \leq \epsilon.$$

Now, using (19), the residual after t -th step is given by:

$$M - U^t (\widehat{V}^{t+1})^\dagger = (I - U^t (U^t)^\dagger) M - U^t F^\dagger.$$

That is,

$$\|M - U^t (\widehat{V}^{t+1})^\dagger\|_F \leq \|(I - U^t (U^t)^\dagger) M\|_F + \|F\|_F \leq \sqrt{k} \|(I - U^t (U^t)^\dagger) U^* \Sigma^*\|_2 + \|F\|_F \leq \sqrt{k} \sigma_1^* \text{dist}(\widehat{U}^t, U^*).$$

Now, using the fact that $\text{dist}(U^t, U^*) \leq \epsilon$ and the above equation, we get:

$$\|M - U^t (\widehat{V}^{t+1})^\dagger\|_F \leq \sqrt{k} \sigma_1^* \epsilon + \|F\|_F \stackrel{\zeta_1}{\leq} \sqrt{k} \sigma_1^* \epsilon + \sigma_1^* \sqrt{k} \epsilon \leq 2\sigma_1^* \sqrt{k} \epsilon,$$

where ζ_1 follows by Lemma 5.6 and setting δ_{2k} appropriately. Theorem 2.5 now follows by setting $\epsilon' = 2\sqrt{k}\|M\|_F \epsilon$. \square

C.1 Initialization: Proofs

Proof Of Lemma 5.2. From Lemma C.1, we see that U^0 obtained after step 3 of Algorithm 2 satisfies: $\text{dist}(U^0, U^*) \leq \frac{1}{64k}$. Lemma now follows by using the above mentioned observation with Lemma C.2. \square

We now provide the two results used in the above lemma.

Lemma C.1. *After step 3 in Algorithm 2, whp we have:*

$$\text{dist}(U^0, U^*) \leq \frac{1}{64k}$$

Proof. From Theorem 3.1 in [12], we have the following result:

$$\|M - M_k\|_2 \leq C \left(\frac{k}{p\sqrt{mn}} \right)^{\frac{1}{2}} \|M\|_F.$$

Let $U^{(0)}\Sigma V^\dagger$ be the top k singular components of M_k . We also have:

$$\begin{aligned} \|M - M_k\|_2^2 &= \left\| U^* \Sigma^* (V^*)^\dagger - U^{(0)} \Sigma V^\dagger \right\|_2^2 \\ &= \left\| U^* \Sigma^* (V^*)^\dagger - U^{(0)} \left(U^{(0)} \right)^\dagger U^* \Sigma^* (V^*)^\dagger + U^{(0)} \left(U^{(0)} \right)^\dagger U^* \Sigma^* (V^*)^\dagger - U^{(0)} \Sigma V^\dagger \right\|_2^2 \\ &= \left\| \left(I - U^{(0)} \left(U^{(0)} \right)^\dagger \right) U^* \Sigma^* (V^*)^\dagger + U^{(0)} \left(\left(U^{(0)} \right)^\dagger U^* \Sigma^* (V^*)^\dagger - \Sigma V^\dagger \right) \right\|_2^2 \\ &\stackrel{(\zeta_1)}{\geq} \left\| \left(I - U^{(0)} \left(U^{(0)} \right)^\dagger \right) U^* \Sigma^* (V^*)^\dagger \right\|_2^2 = \left\| \left(U_\perp^{(0)} \right)^\dagger U^* \Sigma^* \right\|_2^2 \geq (\sigma_k^*)^2 \left\| \left(U_\perp^{(0)} \right)^\dagger U^* \right\|_2^2, \end{aligned}$$

where (ζ_1) follows from the fact that the column space of the first two terms in the equation is $U_\perp^{(0)}$ where as the column space of the last two terms is $U^{(0)}$. Using the above two inequalities, we get:

$$\left\| \left(U_\perp^{(0)} \right)^\dagger U^* \right\|_2 \leq C \cdot \frac{\sigma_1^*}{\sigma_k^*} \cdot \frac{k}{\sqrt{mp}} \leq \frac{1}{10^4 k},$$

if $p > \frac{C' k^4 \log n}{m} \cdot \frac{(\sigma_1^*)^2}{(\sigma_k^*)^2}$ for a large enough constant C' . \square

Lemma C.2. *(Analysis of step 4 of Algorithm 2) Suppose U^* is incoherent with parameter μ and U is an orthonormal column matrix such that $\text{dist}(U, U^*) \leq \frac{1}{64k}$. Let U^c be obtained from U by setting all entries greater than $\frac{2\mu\sqrt{k}}{\sqrt{n}}$ to zero. Let \tilde{U} be an orthonormal basis of U^c . Then,*

- $\text{dist}(\tilde{U}, U^*) \leq 1/2$ and
- \tilde{U} is incoherent with parameter $4\mu\sqrt{k}$.

Proof. Since $\text{dist}(U, U^*) \leq d$, we have that for every i , $\exists \check{u}_i \in \text{Span}(U^*)$, $\|\check{u}_i\|_2 = 1$ such that $\langle u_i, \check{u}_i \rangle \geq \sqrt{1 - d^2}$. Also, since $\check{u}_i \in \text{Span}(U^*)$, we have that \check{u}_i is incoherent with parameter $\mu\sqrt{k}$:

$$\|\check{u}_i\|_2 = 1 \text{ and } \|\check{u}_i\|_\infty \leq \frac{\mu\sqrt{k}}{\sqrt{m}}.$$

Let u_i^c be the vector obtained by setting all the elements of u_i with magnitude greater than $\frac{2\mu\sqrt{k}}{\sqrt{m}}$ to zero and let $u_i^{\bar{c}} \stackrel{\text{def}}{=} u_i - u_i^c$. Now, note that if for element j of u_i we have $|u_i^j| > \frac{2\mu\sqrt{k}}{\sqrt{m}}$, then, $|(u_i^c)^j - \check{u}_i^j| = |\check{u}_i^j| \leq \frac{\mu\sqrt{k}}{\sqrt{m}} \leq |u_i^j - \check{u}_i^j|$. Hence,

$$\|u_i^c - \check{u}_i\|_2 \stackrel{(\zeta_1)}{\leq} \|u_i - \check{u}_i\|_2 = \left(\|u_i\|_2^2 + \|\check{u}_i\|_2^2 - 2\langle u_i, \check{u}_i \rangle \right)^{\frac{1}{2}} \leq \sqrt{2}d,$$

This also implies the following:

$$\begin{aligned} \|u_i^c\|_2 &\geq \|\check{u}_i\|_2 - \sqrt{2}d = 1 - \sqrt{2}d, \text{ and} \\ \|u_i^{\bar{c}}\|_2 &\leq \sqrt{1 - \|u_i^c\|_2^2} \leq \sqrt{2d(\sqrt{2} - d)} \leq 2\sqrt{d}, \text{ for } d < \frac{1}{\sqrt{2}}. \end{aligned}$$

Let $U^c = \tilde{U}\Lambda^{-1}$ (QR decomposition). Then, for any $u_{\perp}^* \in \text{Span}(U_{\perp}^*)$ we have:

$$\begin{aligned} \|(u_{\perp}^*)^{\dagger} \tilde{U}\|_2 &= \|(u_{\perp}^*)^{\dagger} U^c \Lambda\|_2 \leq \|(u_{\perp}^*)^{\dagger} U^c\|_2 \|\Lambda\|_2 \leq \left(\|(u_{\perp}^*)^{\dagger} U\|_2 + \|(u_{\perp}^*)^{\dagger} U^{\bar{c}}\|_2 \right) \|\Lambda\|_2 \\ &\leq (d + \|U^{\bar{c}}\|_2) \|\Lambda\|_2 \leq (d + \|U^{\bar{c}}\|_F) \|\Lambda\|_2 \leq (d + 2\sqrt{kd}) \|\Lambda\|_2 \leq 3\sqrt{kd} \|\Lambda\|_2. \end{aligned}$$

We now bound $\|\Lambda\|_2$ as follows:

$$\|\Lambda\|_2^2 = \frac{1}{\sigma_{\min}(\Lambda^{-1})^2} = \frac{1}{\sigma_{\min}(\tilde{U}\Lambda^{-1})^2} = \frac{1}{\sigma_{\min}(U^c)^2} \leq \frac{1}{1 - \|U^{\bar{c}}\|_2^2} \leq \frac{1}{1 - 4kd} \leq 4/3,$$

where we used the fact that $d < \frac{1}{16k}$. So we have:

$$\|(u_{\perp}^*)^{\dagger} \tilde{U}\|_2 \leq 3\sqrt{kd} \cdot 4/3 = 4\sqrt{kd}.$$

This proves the first part of the lemma.

Incoherence of \tilde{U} follows using the following set of inequalities:

$$\mu(\tilde{U}) = \frac{\sqrt{m}}{\sqrt{k}} \max_i \|e_i^{\dagger} \tilde{U}\|_2 \leq \frac{\sqrt{m}}{\sqrt{k}} \max_i \|e_i^{\dagger} U^c \Lambda\|_2 \leq \frac{\sqrt{m}}{\sqrt{k}} \max_i \|e_i^{\dagger} U^c\|_2 \|\Lambda\|_2 \leq 4\mu\sqrt{k}.$$

□

C.2 Rank-1 Matrix Completion: Proofs

Proof Of Lemma 5.3. Using the definition of spectral norm,

$$\|B^{-1}(\langle u^*, u^t \rangle B - C) v^*\|_2 \leq \|B^{-1}\|_2 \|(\langle u^*, u^t \rangle B - C) v^*\|_2.$$

As B is a diagonal matrix, $\|B^{-1}\|_2 = \frac{1}{\min_i B_{ii}} \leq \frac{1}{1-\delta_2}$, where the last inequality follows using Lemma C.3. The lemma now follows using the above observation and Lemma C.4. □

Lemma C.3. Let $M = \sigma^* u^* (v^*)^{\dagger}$, p , Ω , u^t be as defined in Lemma 5.3. Then, w.p. at least $1 - \frac{1}{n^3}$,

$$\left| \frac{\sum_{i:(i,j) \in \Omega} (u_i^t)^2}{p} - 1 \right| \leq \delta_2, \quad \left| \frac{\sum_{i:(i,j) \in \Omega} u_i^t u_i^*}{p} - \langle u^t, u^* \rangle \right| \leq \delta_2.$$

Proof. Since the first part of the lemma is a direct consequence of the second part, we will prove only the second part. Let δ_{ij} be a Bernoulli random variable that indicates membership of index $(i, j) \in \Omega$. That is, $\delta_{ij} = 1$ w.p. p and 0 otherwise. Define $Z_j = \frac{1}{p} \sum_i \delta_{ij} u_i^t u_i^*$. Note that $\mathbb{E}[Z_j] = \langle u^t, u^* \rangle$. Furthermore, $\mathbb{E}[Z_j^2] = \left(\frac{1}{p} - 1\right) \sum_i (u_i^t u_i^*)^2 \leq \frac{\mu_1^2}{mp}$ and $\max_i |u_i^t u_i^*| \leq \frac{\mu_1^2}{m}$. Using Bernstein's inequality, we get:

$$\Pr(|Z_j - \langle u^t, u^* \rangle| > \delta_2) \leq \exp\left(-\frac{\delta_2^2 mp/2}{\mu_1^2 + \mu_1^2 \delta_2/3}\right). \quad (50)$$

Using union bound (for all j) and for $p \geq \frac{9\mu_1^2 \log n}{m\delta_2^2}$, w.p. $1 - \frac{1}{n^3}$: $\forall j, \langle u^t, u^* \rangle - \delta_2 \leq Z_j \leq \langle u^t, u^* \rangle + \delta_2$. \square

Lemma C.4. Let $M = \sigma^* u^* (v^*)^\dagger$, p , Ω , u^t be as defined in Lemma 5.3. Then, w.p. at least $1 - \frac{1}{n^3}$,

$$\|(\langle u^*, u^t \rangle B - C)v^*\|_2 \leq \delta_2 \sqrt{1 - \langle u^*, u^t \rangle^2}.$$

Proof. Let $x \in \mathbb{R}^n$ be a unit vector. Then, $\forall x$:

$$\begin{aligned} x^\dagger (\langle u^*, u^t \rangle B - C)v^* &= \frac{1}{p} \sum_{ij \in \Omega} x_j v_j^* (\langle u^*, u^t \rangle (u_i^t)^2 - u_i^t u_i^*) \\ &\stackrel{(\zeta_1)}{\leq} \frac{1}{p} C \sqrt{mp} \sqrt{\sum_j x_j^2 (v_j^*)^2} \sqrt{\sum_i (\langle u^*, u^t \rangle (u_i^t)^2 - u_i^t u_i^*)^2}, \\ &\stackrel{(\zeta_2)}{\leq} \frac{1}{p} C \frac{\sqrt{mp} \mu_1^2}{n} \sqrt{1 - \langle u^*, u^t \rangle^2}, \end{aligned} \quad (51)$$

where $C > 0$ is a global constant and (ζ_1) follows by using a modified version of Lemma 6.1 by [12] (see Lemma C.5) and (ζ_2) follows by using incoherence of v^* and u^t . Lemma now follows by observing that $\max_{x, \|x\|_2=1} x^\dagger (\langle u^*, u^t \rangle B - C)v^* = \|(\langle u^*, u^t \rangle B - C)v^*\|_2$ and $p > \frac{C\mu_1^2 \log n}{m\delta_2^2}$. \square

Proof of Lemma 5.4. Using (15) and using the fact that B, C are diagonal matrices:

$$\widehat{v}_j^{t+1} = \sigma^* \langle u^t, u^* \rangle v_j^* - \frac{\sigma^*}{B_{jj}} (\langle u^t, u^* \rangle B_{jj} - C_{jj}) v_j^*.$$

We bound the largest magnitude of elements in \widehat{v}^{t+1} as follows. For every $j \in [n]$, we have:

$$\begin{aligned} |\widehat{v}_j^{t+1}| &\leq |\sigma^* \langle u^t, u^* \rangle v_j^*| + \left| \frac{\sigma^*}{B_{jj}} (\langle u^t, u^* \rangle B_{jj} - C_{jj}) v_j^* \right| \\ &\stackrel{(\zeta_1)}{\leq} \sigma^* \langle u^t, u^* \rangle \frac{\mu}{\sqrt{n}} + \frac{\sigma^*}{1 - \delta_2} (\langle u^t, u^* \rangle (1 + \delta_2) + (|\langle u^t, u^* \rangle| + \delta_2)) \frac{\mu}{\sqrt{n}} \\ &\leq \frac{3\sigma^* (1 + \delta_2) \mu}{1 - \delta_2} \leq \frac{\sigma^* \mu_1}{2\sqrt{n}}, \end{aligned}$$

where (ζ_1) follows from the fact that $1 - \delta_2 \leq B_{jj} \leq 1 + \delta_2$ and $|C_{jj}| \leq (|\langle u^t, u^* \rangle| + \delta_2)$ (please refer Lemma C.3).

Also, from (17) we see that:

$$\begin{aligned}\|\widehat{v}^{t+1}\|_2 &\geq \langle \widehat{v}^{t+1}, v^* \rangle \geq \sigma^* \langle u^t, u^* \rangle - 2\sigma^* \delta_2 \sqrt{1 - \langle u^t, u^* \rangle^2} \\ &\geq \sigma^* \langle u^0, u^* \rangle - 2\sigma^* \delta_2 \sqrt{1 - \langle u^0, u^* \rangle^2} \\ &\stackrel{(\zeta_1)}{\geq} \frac{\sigma^*}{2},\end{aligned}$$

where (ζ_1) follows from the fact that $\text{dist}(u^0, u^*) \leq \frac{3}{50}$ (please refer Lemma 5.2). Using the above two inequalities, we obtain:

$$\|v^{t+1}\|_\infty = \frac{\|\widehat{v}^{t+1}\|_\infty}{\|\widehat{v}^{t+1}\|_2} \leq \frac{\left(\frac{\sigma^* \mu_1}{2\sqrt{n}}\right)}{\left(\frac{\sigma^*}{2}\right)} = \frac{\mu_1}{\sqrt{n}}.$$

This finishes the proof. \square

Lemma C.5 (Modified version of Lemma 6.1 of [12]). *Let Ω be a set of indices sampled uniformly at random from $[m] \times [n]$ with each element of $[m] \times [n]$ sampled independently with probability $p \geq \frac{C \log n}{m}$. Then, w.p. at least $1 - \frac{1}{n^3}$, $\forall x \in \mathbb{R}^m, y \in \mathbb{R}^n$ s.t. $\sum_i x_i = 0$, we have: $\sum_{ij \in \Omega} x_i y_j \leq C\sqrt{\sqrt{mnp}}\|x\|_2\|y\|_2$, where $C > 0$ is a global constant.*

C.3 General Rank- k Matrix Completion: Proofs

Proof of Lemma 5.5. From the decoupled update equation, (20), we obtain:

$$(V^{t+1})^{(j)} = (R^{(t+1)})^{-1}(D^j - (B^j)^{-1}(B^j D^j - C^j))\Sigma^*(V^*)^{(j)}, \quad 1 \leq j \leq n.$$

We bound the two norm of the $(V^{t+1})^{(j)}$ as follows:

$$\begin{aligned}\|(V^{t+1})^{(j)}\|_2 &\leq \frac{\sigma_1 \|(V^*)^{(j)}\|_2}{\sigma_{\min}(R^{(t+1)})} (\|D^j\|_2 + \|(B^j)^{-1}(B^j D^j - C^j)\|_2) \\ &\leq \frac{\sigma_1 \|(V^*)^{(j)}\|_2}{\sigma_{\min}(R^{(t+1)})} \left(\|D^j\|_2 + \frac{\|B^j D^j\|_2 + \|C^j\|_2}{\sigma_{\min}(B^j)} \right) \\ &\stackrel{(\zeta_1)}{\leq} \frac{\sigma_1 \frac{\mu\sqrt{k}}{\sqrt{n}}}{\sigma_k^* \sqrt{1 - \text{dist}^2(U^{(t)}, U^*)} - \frac{\sigma_1^* \delta_{2k} k \text{dist}(U^{(t)}, U^*)}{1 - \delta_{2k}}} \left(1 + \frac{(1 + \delta_{2k}) + (1 + \delta_{2k})}{1 - \delta_{2k}} \right) \\ &\leq \frac{4\sigma_1 \frac{\mu\sqrt{k}}{\sqrt{n}}}{\sigma_k^* \sqrt{1 - \text{dist}^2(U^{(0)}, U^*)} - \frac{\sigma_1^* \delta_{2k} k \text{dist}(U^{(0)}, U^*)}{1 - \delta_{2k}}} \leq \frac{\left(\frac{16\sigma_1^* \mu}{\sigma_k^*}\right) \sqrt{k}}{\sqrt{n}},\end{aligned}$$

where we used the following inequalities in (ζ_1) :

$$\|(V^*)^j\|_2 \leq \frac{\mu\sqrt{k}}{\sqrt{n}}, \tag{52}$$

$$\sigma_{\min}(R^{(t+1)}) \geq \sigma_k^* \sqrt{1 - \text{dist}^2(U^{(t)}, U^*)} - \sigma_1^* \delta_{2k} k \text{dist}(U^{(t)}, U^*), \tag{53}$$

$$\sigma_{\min}(B^j) \geq 1 - \delta_{2k} \text{ and } \sigma_{\max}(B^j) \leq 1 + \delta_{2k}, \tag{54}$$

$$\sigma_{\max}(C^j) \leq 1 + \delta_{2k} \text{ and } \tag{55}$$

$$\sigma_{\max}(D^j) \leq 1, \tag{56}$$

where (52) follows from the incoherence of V^* , (53) follows from an analysis similar to the proof of Lemma 4.7, (54) follows from (the proof of) Lemma C.6, (55) follows from Lemma C.7 and finally (56) follows from the fact that $D^j = (U^t)^\dagger U^*$ with U^t and U^* being orthonormal column matrices. \square

Proof of Lemma 5.6. Note that,

$$\begin{aligned} \|F(\Sigma^*)^{-1}\|_2 &\leq \|F(\Sigma^*)^{-1}\|_F = \|B^{-1}(BD - C)v^*\|_2 \\ &\leq \|B^{-1}\|_2 \|(BD - C)v^*\|_2 \\ &\leq \frac{\delta_{2k}}{1 - \delta_{2k}} \text{dist}(U^t, U^*), \end{aligned} \quad (57)$$

where the last inequality follows using Lemma C.6 and Lemma C.8. \square

We now bound $\|B^{-1}\|_2$ and $\|C^j\|_2$, which is required by our bound for F as well as for our incoherence proof.

Lemma C.6. *Let M, Ω, p , and U^t be as defined in Theorem 2.5 and Lemma 5.6. Then, w.p. at least $1 - \frac{1}{n^3}$:*

$$\|B^{-1}\|_2 \leq \frac{1}{1 - \delta_{2k}}. \quad (58)$$

Proof of Lemma C.6. We have:

$$\|B^{-1}\|_2 = \frac{1}{\sigma_{\min}(B)} = \frac{1}{\min_{x, \|x\|=1} x^\dagger Bx},$$

where $x \in \mathbb{R}^{nk}$. Let $x = \text{vec}(X)$, i.e., x_p is the p -th column of X and x^j is the j -th row of X . Now, $\forall x$,

$$x^\dagger Bx = \sum_j (x^j)^\dagger B^j (x^j) \geq \min_j \sigma_{\min}(B^j).$$

Lemma would follow using the bound on $\sigma_{\min}(B^j), \forall j$ that we show below.

Lower bound on $\sigma_{\min}(B^j)$: Consider any $w \in \mathbb{R}^k$ such that $\|w\|_2 = 1$. We have:

$$Z = w^\dagger B^j w = \frac{1}{p} \sum_{i:(i,j) \in \Omega} \langle w, (U^t)^{(i)} \rangle^2 = \frac{1}{p} \sum_i \delta_{ij} \langle w, (U^t)^{(i)} \rangle^2.$$

Note that, $\mathbb{E}[Z] = w^\dagger U U^\dagger w = w^\dagger w = 1$ and $\mathbb{E}[Z^2] = \frac{1}{p} \sum_i \langle w, (U^t)^{(i)} \rangle^4 \leq \frac{\mu_1^2 k}{mp} \sum_i \langle w, (U^t)^{(i)} \rangle^2 = \frac{\mu_1^2 k}{mp}$, where the second last inequality follows using incoherence of U^t . Similarly, $\max_i |\langle w, (U^t)^{(i)} \rangle^2| \leq \frac{\mu_1^2 k}{mp}$. Hence, using Bernstein's inequality:

$$\Pr(|Z - \mathbb{E}[Z]| \geq \delta_{2k}) \leq \exp\left(-\frac{\delta_{2k}^2/2}{1 + \delta_{2k}/3} \frac{mp}{\mu_1^2 k}\right).$$

That is, by using p as in the statement of the lemma with the above equation and using union bound, we get (w.p. $> 1 - 1/n^3$): $\forall w, j \quad w^\dagger B^j w \geq 1 - \delta_{2k}$. That is, $\forall j, \sigma_{\min}(B^j) \geq (1 - \delta_{2k})$. \square

Lemma C.7. Let M, Ω, p , and U^t be as defined in Theorem 2.5 and Lemma 5.6. Also, let $C^j \in \mathbb{R}^{k \times k}$ be defined as: $C^j = \frac{1}{p} \sum_{i:(i,j) \in \Omega} (U^t)^{(i)} (U^*)^{(i)\dagger}$. Then, w.p. at least $1 - \frac{1}{n^3}$:

$$\|C^j\|_2 \leq 1 + \delta_{2k}, \forall j \quad (59)$$

Proof of Lemma C.7. Let $x \in \mathbb{R}^k$ and $y \in \mathbb{R}^k$ be two arbitrary unit vectors. Then,

$$x^T C^j y = \frac{1}{p} \sum_{i:(i,j) \in \Omega} (x^\dagger (U^t)^{(i)}) (y^\dagger (U^*)^{(i)}).$$

That is, $Z = x^T C^j y = \frac{1}{p} \sum_i \delta_{ij} (x^\dagger (U^t)^{(i)}) (y^\dagger (U^*)^{(i)})$. Note that, $\mathbb{E}[Z] = x^\dagger (U^t)^\dagger U^* y$, $\mathbb{E}[Z^2] = \frac{1}{p} \sum_i (x^\dagger (U^t)^{(i)})^2 (y^\dagger (U^*)^{(i)})^2 \leq \frac{\mu^2}{mp} x^\dagger (U^t)^\dagger U^t x = \frac{\mu^2 k}{mp}$ and $\max_i |(x^\dagger (U^t)^{(i)}) (y^\dagger (U^*)^{(i)})| \leq \frac{\mu^2 k}{m}$. Lemma now follows using Bernstein's inequality and using bound for p given in the lemma statement. \square

Finally, we provide a lemma to bound the second part of the error term (F).

Lemma C.8. Let M, Ω, p , and U^t be as defined in Theorem 2.5 and Lemma 5.6. Then, w.p. at least $1 - \frac{1}{n^3}$:

$$\|(BD - C)v^*\|_2 \leq \delta_{2k} \text{dist}(V^{t+1}, V^*), \quad (60)$$

where $v^* = \text{vec}(V^*)$, i.e. $v^* = \begin{bmatrix} V_1^* \\ \vdots \\ V_k^* \end{bmatrix}$.

Proof of Lemma C.8. Let $X \in \mathbb{R}^{n \times k}$ and let $x = \text{vec}(X) \in \mathbb{R}^{nk}$ s.t. $\|x\|_2 = 1$. Also, let x_p be the p -th column of X and x^j be the j -th column of X .

Let $u^i = (U^t)^{(i)}$ and $u^{*(i)} = (U^*)^{(i)}$. Also, let $H^j = (B^j D - C^j)$, i.e.,

$$H^j = \frac{1}{p} \sum_{i:(i,j) \in \Omega} u^i (u^i)^\dagger (U^t)^\dagger U^* - u^i (u^{*(i)})^\dagger = \frac{1}{p} \sum_{i:(i,j) \in \Omega} H_i^j,$$

where $H_i^j \in \mathbb{R}^{k \times k}$. Note that,

$$\sum_i H_i^j = (U^t)^\dagger U^t (U^t)^\dagger U^* - (U^t)^\dagger U^* = 0. \quad (61)$$

Now, $x^\dagger (BD - C)v^* = \sum_j (x^j)^\dagger (B^j D - C^j)(V^*)^{(j)} = \frac{1}{p} \sum_{pq} \sum_{(i,j) \in \Omega} x_p^j V_{jq}^* (H_i^j)_{pq}$. Also, using (61), $\forall(p, q)$:

$$\sum_i (H_i^j)_{pq} = 0.$$

Hence, applying Lemma C.5, we get w.p. at least $1 - \frac{1}{n^3}$:

$$x^\dagger (BD - C)v^* = \sum_j (x^j)^\dagger (B^j D - C^j)(V^*)^{(j)} \leq \frac{1}{p} \sum_{pq} \sqrt{\sum_j (x_p^j)^2 (V_{jq}^*)^2} \sqrt{\sum_i (H_i^j)_{pq}^2}. \quad (62)$$

Also,

$$\begin{aligned}
\sum_i (H_i^j)_{pq}^2 &= \sum_i (u_p^i)^2 ((u^i)^\dagger (U^t)^\dagger U_q^* - U_{iq}^*)^2 \leq \max_i (u_p^i)^2 \sum_i ((u^i)^\dagger (U^t)^\dagger U_q^* - U_{iq}^*)^2 \\
&= \max_i (u_p^i)^2 (1 - \|U^t U_q^*\|_2^2) \leq \frac{\mu_1^2 k}{m} \text{dist}(U^t, U^*)^2.
\end{aligned} \tag{63}$$

Using (62), (63) and incoherence of V^* , we get (w.p. $1 - 1/n^3$), $\forall x$:

$$x^\dagger (BD - C)v^* \leq \sum_{pq} \frac{\mu_1^2 k}{mp} \text{dist}(U^t, U^*) \|x_p\|_2 \leq \delta_{2k} \text{dist}(U^t, U^*),$$

where we used the fact that $\sum_p \|x_p\|_2 \leq \sqrt{k} \|x\|_2 = \sqrt{k}$ in the last step. Lemma now follows by observing $\max_{x, \|x\|=1} x^\dagger (BD - C)v^* = \|(BD - C)v^*\|_2$. \square